1990

Perceptual aspects of a three-dimensional vowel space

John Warren Hawks

WASHINGTON UNIVERSITY
Department of Speech and Hearing
Program in Communication Sciences

Dissertation Committee:
James D. Miller, Chairman
Ira J. Hirsh
A. Maynard Engebretson
Margaritis S. Fourakis

Perceptual Aspects of a Three-dimensional Vowel Space

by

John Warren Hawks

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August, 1990
Saint Louis, Missouri

Abstract

Chairman: James D. Miller

One of the methods that has been utilized for vowel classification in natural speech posits a three-dimensional space, defined by certain acoustic characteristics of the vowels related to the first three significant prominences, or formants, in their short-term spectra ($F1$, $F2$, and $F3$) and voice pitch ($F0$). Within this space, productions of like vowels are mapped onto subspaces, called target zones. Here, the characteristics of this space and its subspaces were studied by means of subject responses to synthetic vowel-like tokens representing unique points in this space. The first experiment utilized the identifications of eight listeners to construct a vowel map of this space. Vowel categories for American English can be represented as abutting and non-overlapping target zones which correctly classify over 99% of the plurality-based identifications. The first two formants ($F1$ and $F2$) appeared to be the primary determinants in the identification of non-retroflex vowels. The third formant ($F3$) determined the perception of retroflex (r-colored) vowels, and contributed to the phonetic saliency of other vowel categories. Furthermore, phonetic saliency varied in an orderly way with location within a vowel target zone.

A second experiment investigated the discrimination of complex sounds represented as points in the three-dimensional vowel space by estimating difference limens (DLs), expressed as distance in the space, for synthetic, vowel-like tokens. Four subjects were used to estimate DLs along 102 straight-line continua. Continua emanated in six directions from 17 locations in the space. Movement along continua resulted in distinct, multi-formant patterns of frequency change which varied with the direction and axis of movement. The average DL for distance across all continua was estimated to be .01 log units, but DLs were significantly different for the three axes of the space. Considerable variation found in DLs associated

with individual reference points may be related to differences in reference formant patterns. The DL results for multiple-formant-change continua were found to be significantly smaller than similar DL estimates for single-formant-change continua. Many of these differences could be accounted for by an additive model whereby each changing formant contributes independent information to perception.

Acknowledgements

Many people have played many roles in aiding and abetting this dissertation. To anyone of these people whom I unintentionally omit from these acknowledgements, I thank you. I would first and foremost like to thank my wife for her ever-enduring tolerance of this project, the many extra hours it took me away from her, the completion postponements, and my temporary insanity. I only hope I can repay this debt. I also thank my children, Ben, Adam, and the little one in the "oven" for constantly reminding me of what life is really about. I hope that now I can start living it with you more fully.

I would like to thank and acknowledge my dissertation committee, Drs. James D. Miller, Ira Hirsh, Maynard Engebretson, and Marios Fourakis, for their help, advice, and encouragement. Dr. Miller, my advisor, deserves special thanks in that he has been instrumental in shaping my education and career since first meeting with me in 1982 to discuss graduate school. He has led me, advised me, employed me, and taught me much about the right and the wrong ways to do things in this business. Special thanks also go to my best friend and officemate, Marios Fourakis for all the hours spent sharing what he knows and keeping me on track.

Although I am pleased that most of the computer programming required of this project I was able to do for myself, thanks go to Steven J. Sadoff and Frank Kramer for their invaluable assistance and to the late Dennis Klatt for providing the world with a wonderful tool in his software synthesizer. I am extremely grateful for the assistance and discussion on statistical issues from Caroline B. Monahan, Martha Storandt, and Janet Weisenberger, and on psychophysical issues from Bob Gilkey. Thanks also go to the CID technical support staff, headed by Arnold Heidbreder, for their ongoing efforts to ensure that no project is short-circuited.

A special acknowledgement to my fellow doctoral students, Chip Nicholas, Punita Singh, Steve Sadoff, and Lyn Shields, for their comradery during this whole ordeal. Thanks also to all those who served as subjects for the pilot work and the experiments for their countless hours of listening, as well as to another officemate, Dr. Michael Gottfried, for numerous discussions and encouragement.

This work was supported by a grant from NIDCD.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and background

1.1 Introduction

A common goal of all theories of speech perception is (or should be) to explain how a listener converts the acoustic speech signal produced by a talker into a sequence of meaningful linguistic units. According to McKay (1956), theories describing the process that takes place during this conversion can be separated into two categories. Active theories view the listener as an 'active' participant in the speech-perception process and link that process with a knowledge of speech production. In these theories, the conversion is accomplished through reference to internal representations of the motor commands that produced the speech sounds encoded in the signal. Passive theories place the listener in a more passive perceptual role and consider the speech-perception process to be more sensory in nature. Here perception of the acoustic signal is generally the result of directly "decoding" it, after some amount of processing, into meaningful informational elements.

This distinction notwithstanding, Miller (1984a) suggests that all current speech-perception theories may be described by means of a generic three-stage model. The model considers only a "bottom-up" processing sequence, that is, perceiving speech without the aid of additional cues from knowledge of the language, context, talker, environment, or other "top-down" influences on speech processing. Although most theories employ top-down processing, many consider it as necessary only for ambiguous situations or instances when communication is difficult.

In stage 1, some form of spectral analysis is performed on the acoustic speech waveform, yielding information about the distribution of acoustic energy over time in terms of its frequency and intensity. In stage 2, the spectral information derived in stage 1 is transformed into perceptual dimensions. Finally, in stage 3, these perceptual dimensions are in turn transformed into linguistically significant elements, such as phonemes, syllables, or words. While most speech theorists agree that in stage 1 the ear acts as a filter bank, performing continuous short-term spectral analyses on the incoming waveform often simulated by Fourier or linear-prediction analyses, the theoretical descriptions of stages 2 and 3 are quite diverse. Active theories such as the motor theory (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman and Mattingly, 1985) and the analysis-by-synthesis theory (Stevens and Halle, 1967) suggest that the perceptual dimensions of stage 2 are articulatory gestures and, by means of special speech decoders or matching procedures, phonemes are perceptually elicited in stage 3. Some passive theorists (Fant, 1967; Morton and Broadbent, 1967) have proposed that auditory features from stage 1 are mapped onto linguistic-phonetic features in stage 2, leading to the perception of phones in stage 3. Klatt (1979) suggested another approach whereby spectral-envelope patterns derived in stage 1 are submitted to a matching procedure with stored templates of diphone sequences in stage 2. Successful matches are then subjected to further processing with word recognition achieved in stage 3.

More recently, another passive theory has been proposed by Miller (1984). The auditory-perceptual theory (*APT*) of speech perception considers that spectral information from stage 1 elicits a sensory representation by activating sensory pointers in a phonetically relevant three-dimensional perceptual space. The sensory pointers move in the space as new spectral information is processed. A perceptual response results from an integrative-predictive process based largely on the dynamics of the sensory pointers' movement in stage 2. A phonetic-linguistic representation is achieved in stage 3 through the activation of perceptual target zones (*PTZ*) in the auditory-perceptual space by the dynamics of the perceptual response. These *PTZs* correspond to the phones of a language and, upon their activation, a neural symbol code is issued. A fourth, lexical-access stage, is required in this model to find words. It is the auditory-perceptual theory of speech perception in general

and the concept and validity of perceptual target zones for vowels in particular which will provide the framework for the following dissertation.

1.2 The Auditory-Perceptual Theory (*APT*)

The fundamental concepts of the *APT* have been most recently detailed in Miller (1989). Since a basic understanding of these concepts is central to an appreciation of the work to be described, a brief description of those concepts pertaining to perceptual target zones and non-nasalized vowels follows.

1.2.1 Auditory-perceptual space (*APS*)

The auditory-perceptual space (*APS*) is defined in three dimensions, x, y, and z, where

$$\begin{aligned} x &= log(SF3/SF2), \\ y &= log(SF1/SR), \quad and \\ z &= log(SF2/SF1). \end{aligned} \tag{1.1}$$

$SF1$, $SF2$, and $SF3$ correspond to the center frequencies in Hz of the first three significant prominences in the acoustic spectra derived from the short-term spectral analysis of the incoming speech signal, commonly termed formants. SR is a low-frequency reference, called the sensory reference, based on the current talker's vocal characteristics. The sensory reference acts as a normalizing anchor point with an initial value of 168 which is shifted up or down depending on the talker's vocal characteristics. SR is often calculated by the equation,

$$SR = 168(GMF0/168)^{1/3}, \tag{1.2}$$

where $GMF0$ represents the geometric mean of the current talker's fundamental frequency.

Spectral shapes are represented in the *APS* by means of sensory pointers. The location of a sensory pointer is determined by the coordinate values derived from a short-term spectral analysis of a time-windowed segment of the incoming acoustic waveform and is continually updated with new analysis information at regular intervals. We have found that a 24-ms

window shifted in 1-ms steps appears adequate to provide the necessary information. A sensory path is created as the sensory pointer moves through the space, directed by the continually updated spectral information. For our purposes, only glottal-source spectra, i.e., where the sound source is at the glottis and *SF*1 is present, and their associated pointer (*GSSP*) will be considered. The glottal-source sensory pointer, or *GSSP*, is activated, indicating a sensory response, whenever the analysis process detects glottal-source sound above an auditory threshold.

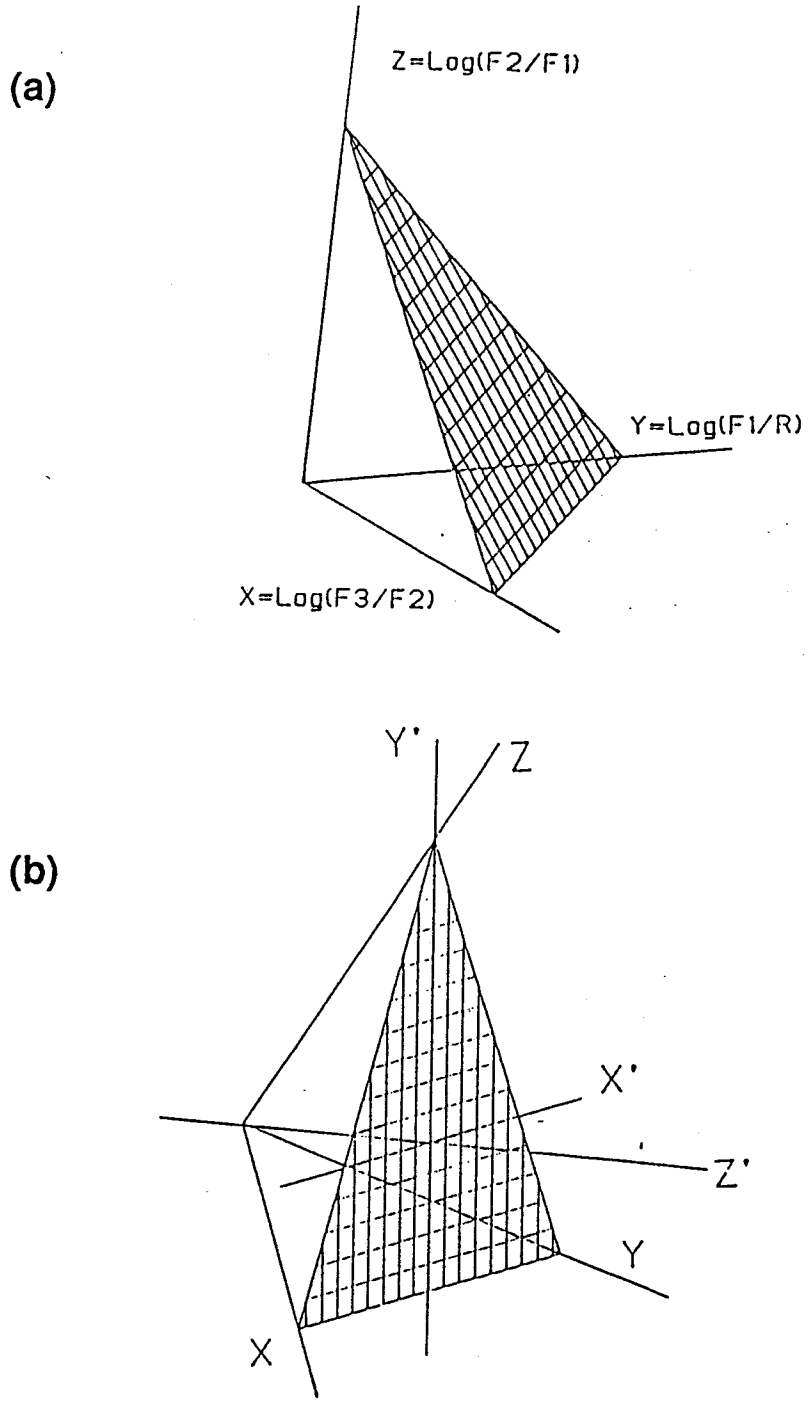
The movements by the sensory pointers are integrated into a unitary perceptual response through a sensory-perceptual transformation represented by a perceptual pointer (*PP*). The location of the perceptual pointer is controlled by the sensory pointers and a neutral point. This concept can be mathematically modeled by considering the perceptual pointer as being attached to the sensory pointers by springs such that movements by the sensory pointers result in movement of the perceptual pointer. Once again, for our purposes, only the glottal-source sensory pointer will elicit the movement of the perceptual pointer. Additionally, the perceptual pointer's movement is adjusted for changes in loudness and goodness. The perceptual pointer does not disappear when the sensory pointers are turned off, but rather, its loudness decays over 100-200 msec as it migrates to a neutral point in the *APS*.

The perceptual pointer may be continually moving through the *APS*, tracing what we call a perceptual path. Since this perceptual path reflects a continuously changing auditory experience, it is posited that the perceptual pointer performs one or a combination of several possible segmentation maneuvers to divide the continuous flow into discrete, auditory-perceptual events. The segmentation maneuvers under consideration include 1) a period of low velocity, 2) an abrupt deceleration, and 3) a high curvature in the perceptual path.

1.2.2 Formant location in the *APS*

When natural vowels are represented as points in the *APS*, the points fall within a "slab" in the *APS* often referred to as the "vowel slab," approximately located in the plane defined by $(x + y + z) = 1.18$ as shown in Figure 1-1a. As an aid to visualization, a simple rotation

Figure 1-1: (a) View of vowel "slab" (gridded area) in *APS* dimensions; (b) view of the same plane in transformed (x', y', z') dimensions.



of the *APS* axes permits a vertical orientation (Figure 1-1b) of this plane where,

$$\begin{aligned} x' &= .70711(y - x), \\ y' &= .8162(z) - .4081(x + y), \textit{and} \\ z' &= .5772(x + y + z). \end{aligned} \tag{1.3}$$

These coordinates are often referred to as "slab" coordinates and views of the *APS* utilizing these coordinates are often used for visual examination and graphical illustration of the *PTZs* for vowels and associated data.

Although the basic x, y, z coordinates of the *APS* are simple log ratios of familiar acoustic speech variables, the transformation of these coordinates to x', y', z' is somewhat more complex and makes the intuitive interpretation of formant location in terms of distance and direction in this coordinate system more difficult. In an effort to reduce this difficulty, the reader is referred to Appendix A for a graphical demonstration and discussion of how certain formant patterns manifest themselves in $x'y'z'$-space.

1.2.3 Concept and estimation of perceptual target zones (*PTZs*)

A perceptual target zone (*PTZ*) is a three-dimensional object or subspace located within the *APS*. The final stage of Miller's generic model, the perception of a phonetic representation, is thought to be accomplished through the activation of a perceptual target zone by way of a segmentation maneuver performed by the perceptual pointer within the defined boundaries of the zone. When a *PTZ* is activated, it is then said to issue a phonetic code or "neural symbol" corresponding to an allophone of the language in question. Current estimates for the shapes and locations of the perceptual target zones for the non-retroflex, monophthongal vowels of American English suggest that they have irregularly shaped, non-overlapping, and abutting boundaries.

Two approaches have been considered for estimating the locations of target zones and their boundaries with each approach having its own list of problems and advantages. In the first approach, locations of the perceptual target zones for the non-retroflex, non-diphthongized vowels of American English have been estimated on the basis of measurements of vowel productions collected from various data sources (Miller, 1987b; 1987c; Miller and

Hawks, 1986; Fourakis and Miller, 1987). These data sources include past studies from the literature as well as measurements made in our own laboratory. A front "slab" view of estimations for these zones[1] in the *APS* based on 435 such measurements can be seen in Figure 1-2, and another more recent and detailed estimation based on 2051 data points in Figure 1-3. Comparison of these figures demonstrates that as additional data points are collected, considerably more intricate boundaries seem to be required to minimize overlap between the zones. Note that the graphic visualizations in Figures 1-2 and 1-3 are two-dimensional, showing only a "front view" of the target zones by x', y' coordinates. To visualize the third dimension, a "side view" perspective is utilized with y', z' coordinates (Figure 1-4), demonstrating irregularly shaped boundaries for the earlier target zones seen in Figure 1-2 in this dimension as well. Figure 1-5 shows this perspective for the most recent target zone estimations (Figure 1-3). Note that due to technical difficulties in delineating concise boundaries in the z' dimension, these zones have been constructed with straight-line boundaries in this dimension, although in theory these boundaries are also assumed to be irregularly shaped. Thus the target zone boundaries shown in Figure 1-2 are somewhat "generalized", since, when fully detailed, these boundaries should vary relative to the z' dimension.

Certain problems exist with using measurements from natural data as the basis for zone construction. First is the problem of phonetic labeling. While all the speech data used in estimating target zones has undergone some type of phonetic verification and can be taken as having perceptual significance, it must be assumed that the vowel tokens represented from the literature were perceived as the phones they were claimed to be representative of. Various levels of phonetic verification are employed ranging from simply what the talker intended to say, to experimenter- and group-verified identifications. Instructions for phonetic identification also vary in the number of phonetic categories provided to choose from. Additional problems with this approach are found in the selection of spectra and the specification of pitch. Picking a representative spectrum for a given vowel utterance can be as arbitrary or subjective a criterion as the experimenter deems appropriate, and additional

[1]The ARPAbet symbol system (Lee and Shoup, 1980) will be used exclusively throughout this thesis for notating phonetic categories (See Table 1.1).

Table 1.1: ARPAbet symbols for representing the phoneme-like units of English within a computer. (From Lee and Shoup, 1980)

Phoneme	Computer Representation		Example	Phoneme	Computer Representation		Example
	1-Character	2-Characters			1-Character	2-Characters	
i	i	IY	beat	p	p	P	pet
ɪ	I	IH	bit	t	t	T	ten
e	e	EY	bait	k	k	K	kit
ɛ	E	EH	bet	b	b	B	bet
æ	@	AE	bat	d	d	D	debt
ɑ	a	AA	Bob	g	g	G	get
ʌ	A	AH	but	h	h	HH	hat
ɔ	c	AO	bought	f	f	F	fat
o	o	OW	boat	θ	T	TH	thing
U	U	UH	book	s	s	S	sat
u	u	UW	boot	š or ʃ	S	SH	shut
ə	x	AX	about	v	v	V	vat
ɨ	X	IX	roses	ð	D	DH	that
ɝ	R	ER	bird	z	z	Z	zoo
aU or aw	W	AW	down	ž or ʒ	Z	ZH	azure
aI or ay	Y	AY	buy	č	C	CH	church
ɔI or ɔy	O	OY	boy	ǰ	J	JH	judge
y	y	Y	you	ʍ	H	WH	which
w	w	W	wit	syl l, l̩	L	EL	battle
r	r	R	rent	syl m, m̩	M	EM	bottom
l	l	L	let	syl n, n̩	N	EN	button
m	m	M	met	flapped t,ɾ	F	DX	batter
n	n	N	net	glottal stop, ʔ	Q	Q	
ŋ	G	NX	sing	Silence	-	-	
				non-speech Segment	!	!	laugh, etc.

AUXILIARY SYMBOLS (1- AND 2-CHARACTER CODES ARE IDENTICAL)

Symbol	Meaning	Symbol	Meaning
+	Morpheme boundary	:3 or .	Fall-rise or non-term juncture
/	Word boundary	* **	Comment (anything except * or **)
#	Utterance boundary	' '	Apos.-surround special symbol in comment
:	Tone group boundary		
		( )	Phoneme class information
:1 or .	Falling or decl. juncture	< >	Phonetic or allophonic escape
:2 or ?	Rising or inter. juncture		

STRESS REPRESENTATIONS (IF PRESENT, MUST IMMEDIATELY FOLLOW THE VOWEL)

Value	Stress Assignment	Value	Stress Assignment
0	No stress	3	Tertiary stress
1	Primary stress	ˈ	(Etc.)
2	Secondary Stress	ˌ	
		ː	

Figure 1-2: Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 435 vowels from natural speech.

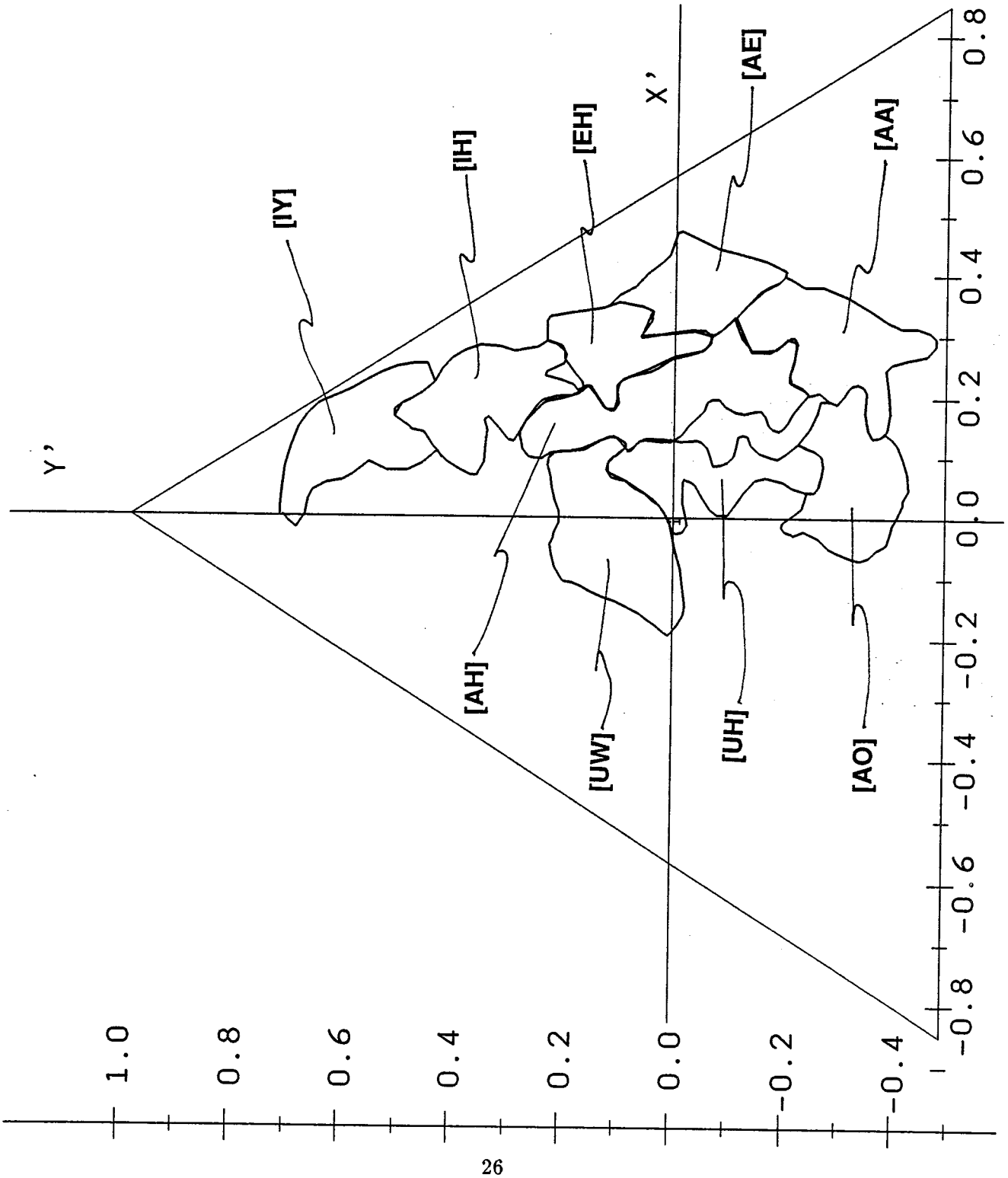


Figure 1-3: Estimations of perceptual target zones for American English in APS $x'y'$ coordinates based on measurements of 2051 vowels from natural speech.

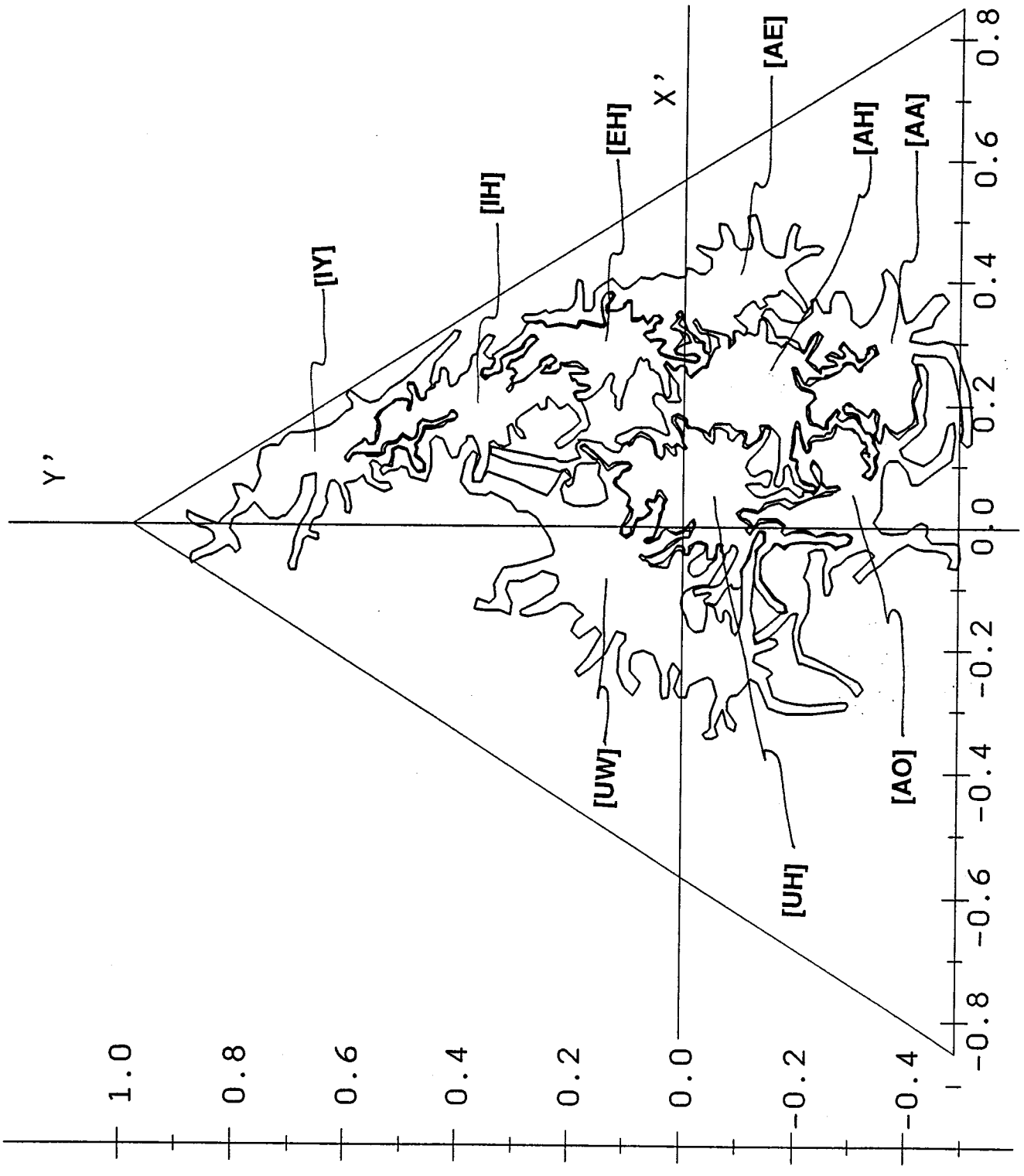


Figure 1-4: Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 435 vowels from natural speech.

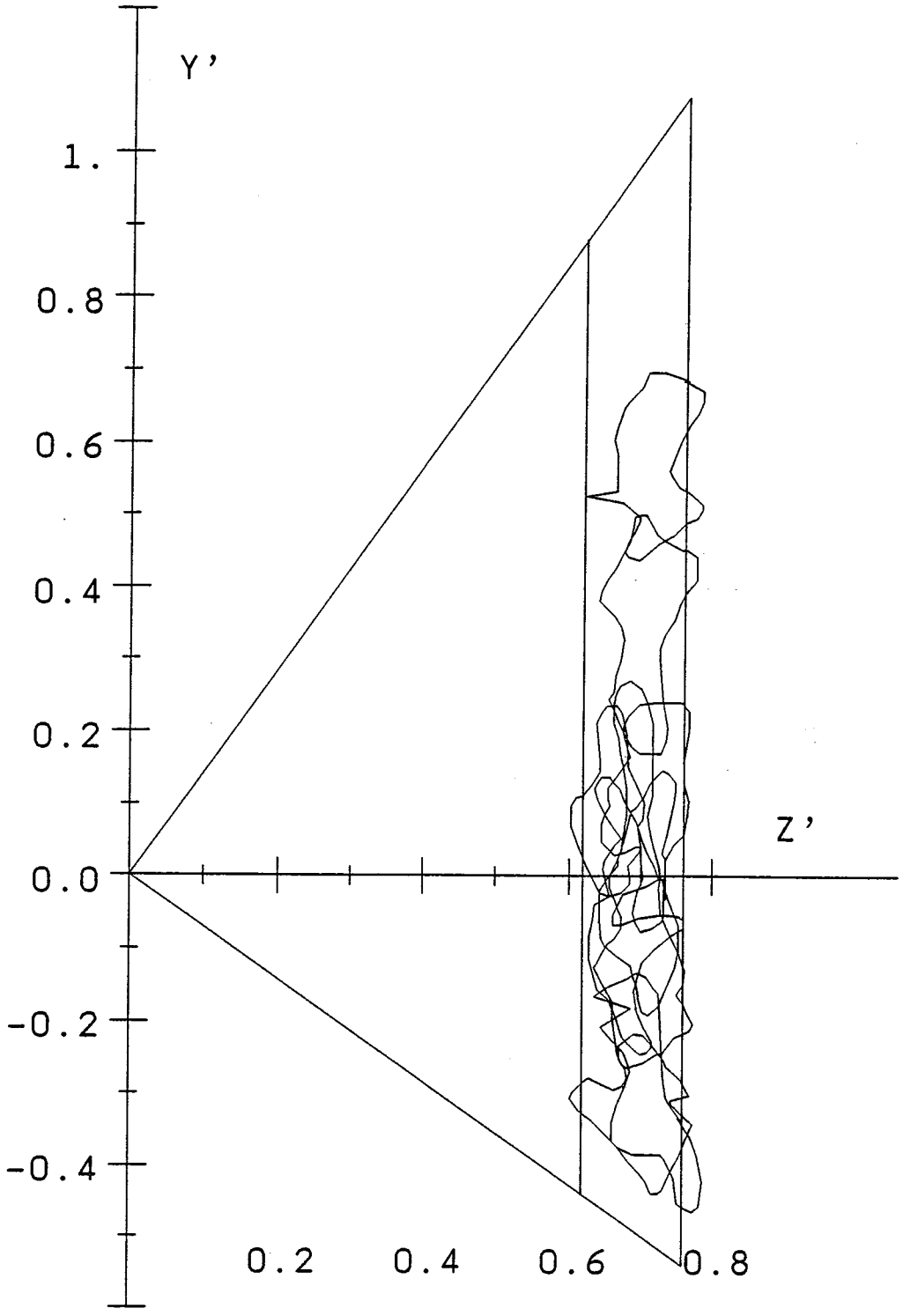
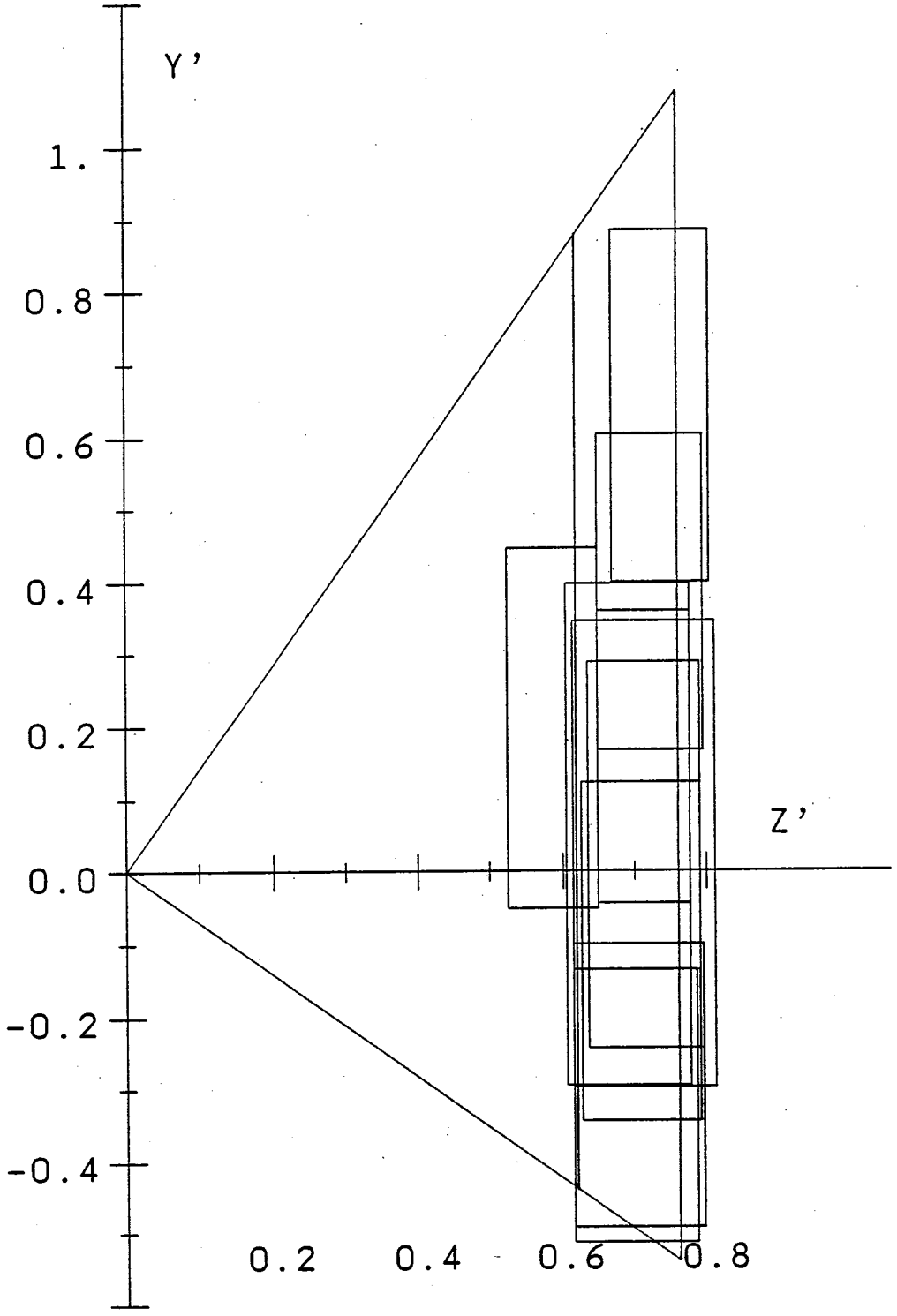


Figure 1-5: Estimations of perceptual target zones for American English in APS $y'z'$ coordinates based on measurements of 2051 vowels from natural speech.



variance can stem from the different measurement techniques which have been employed through the years.

It is currently hypothesized that, given a sufficient number of data points gathered from more talkers and more vowel contexts, the *PTZ* boundaries will stabilize, although there is neither guarantee that this will happen nor an estimation as to the number of points, talkers, or contexts this stabilization may require. This uncertainty questions the validity of basing perceptual target zones solely on such an approach.

The second approach to determining the locations of perceptual target zones utilizes the phonetic identifications of synthetic speech as a basis for mapping the *APS* vowel space. The speech synthesizer is a powerful and often-used scientific tool for investigating hypotheses about complex sounds and has proven particularly useful in speech and auditory research. The synthetic-speech approach has the advantage that, unlike the natural- speech approach which must rely on chance for phonetic identification of locations in *APS*, specific locations in the *APS* can be synthesized and phonetically identified, providing a more complete view of perceptual target zones and their boundaries.

However, the synthetic-speech approach is not without its problems. Since all the variations for the parameters comprising natural speech that elicit perceptual differences in vowel quality have yet to be specified, it is impossible to synthesize vowels which sound completely natural. Furthermore, while it is safe to assume that all these parameters are present in the vowel utterances utilized in the natural speech approach, it can only be speculated that the physical attributes necessary for perceptual salience of synthetic vowels are present. Additionally, the parameters selected to be varied must be questioned as to whether they are correct and sufficient for changing perceived vowel quality. Thus, the validity of using this approach alone is also questioned.

Given consideration of these two approaches for estimation of perceptual target zones and their boundaries, it becomes apparent that neither is clearly superior and that the best solution may be to utilize both approaches with the hope that the results will converge and verify one another.

1.3 Overview of experiments

For decades speech researchers have attempted to find a method or metric based on acoustic parameters whereby the vowels of a language could be described with unique and independent categories, despite differences in talkers, speaking rate, stress, and context. The concept of the perceptual target zone offers a potential answer to this question and provides as well a major underlying foundation to the auditory-perceptual theory. Thus, the validation of this concept is critical not only to the ongoing research efforts within the theory's present framework, but also to the general acceptance of such a theory by the scientific community.

Two experiments will be outlined here aimed at exploring the validity of perceptual target zones for vowels. Experiment I will investigate, by means of phonetic identifications and confidence ratings, listeners' perceptions of simple isolated vowel sounds which have been synthesized from variables provided by distinct points in the *APS*. The results of this experiment will provide a phonetic map of the areas in and around the vowel slab and offer a basis for comparison between this and other vowel classification approaches, including the current *PTZ* estimates. The results of Experiment I will also provide information about individual perceptual differences and mapping reliability and serve as guidelines for Experiment II.

Experiment II will investigate the *APS* at very fine levels of resolution. Indeed, the goal of Experiment II is to estimate the difference limen for various locations in the *APS*. An adaptive up-down procedure utilizing a cued, two-alternative, forced-choice (2AFC) task will be employed. The results of this experiment provide information about the resolution necessary for exploring more precise *PTZ* boundaries, as well as address questions concerning the potential differences in discrimination sensitivity within and between *PTZs*. Additionally, evidence will be presented demonstrating how discrimination of vowel sounds is affected by multiple simultaneous formant changes.

Chapter 2

Experiment I: Perceptual Mapping of the APS Vowel Space.

2.1 Introduction

The purpose of Experiment I is to gather data which can be used to estimate the sizes, shapes, and locations of target zones for the vowels of American English in the *APS* and additionally, to attempt to validate the concept that such zones are non-overlapping, enabling their use for vowel classification. The target zone estimates will be made by means of listeners' identifications, and corresponding confidence ratings, of synthetic tokens which represent specific locations in the auditory-perceptual space. The experiment is motivated by the fact that inspection of estimates for these zones based on the presently available production data (Figures 1-9 and 1-10) shows that portions of the space in the vowel slab remain unaccounted for. Several reasons may account for this. 1) As was previously discussed, the current database may be too limited in terms of an adequate sample of subjects and phonetic environments to modify the current boundaries further. This may be particularly true for unaccounted-for spaces between the *PTZs* and the exterior edges of boundaries. 2) Vocalic sounds corresponding to points in the unaccounted-for spaces may represent speech sounds or portions of speech sounds other than vowels. There is reason to believe that the *PTZs* for [L,R,W,Y] may occupy portions of the vowel slab, as well as the voiced portions of the fricatives [Z,V,DH,ZH,JH], flaps [DX], and voice bars (Miller and Hawks, 1986). 3) Vocalic

sounds corresponding to points in the unaccounted-for spaces may represent speech sounds from languages other than American English. In particular, rounded front vowels, such as those found in German and Swedish may uniquely occupy some of the unaccounted-for space (Jongman, Fourakis, and Sereno, 1989). 4) Vocalic sounds corresponding to points in the unaccounted-for spaces may be perceived as belonging to specific phoneme categories, but are not generally realizable by the human vocal apparatus. Preliminary work indicates that some synthetic vowel sounds utilizing extreme formant values or unlikely relations between formant values can be perceived as English vowels. 5) Vocalic sounds corresponding to points in the unaccounted-for spaces may sound completely unspeechlike. Computation of the values of $SF1$, $SF2$, and $SF3$ for a given value of SR in the spaces currently unaccounted for reveal that some of these points may not be physiologically realizable by the human vocal tract, but should be physically realizable with a digital formant synthesizer (i.e., $F0 < F1 < F2 < F3$). Given the primary purposes and exploratory nature of this experiment, all of these possible reasons cannot be addressed and thus become part of the goals for future research. Delineation of non-vowels will not be attempted and identification of vowels from languages other than American English will be dealt with in only a qualitative manner. Some attempt will be made, however, to limit tokens to those which potentially may be produced by a male voal tract.

Experiment I is additionally motivated by the fact that no studies to our knowledge have investigated an extensive range of sounds that humans are able to produce or perceive as vowels. Many perceptual studies of vowel spaces and their boundaries have been based on the identification of natural speech (Potter and Peterson, 1948; Peterson and Barney, 1952; Fairbanks and Grubb, 1961; Pols, van der Kamp, and Plomp, 1969), and in these experiments an exact description of the stimuli is often difficult to obtain. Of the studies dealing with the perception of synthetic stimuli, most have utilized straight-line (relative to an F1-F2 plot) continua which cross through several vowel spaces (Fry, Abramson, Eimas, and Liberman, 1962; Stevens, Liberman, Studdert-Kennedy, and Öhman, 1969; Repp, Healey, and Crowder, 1979).

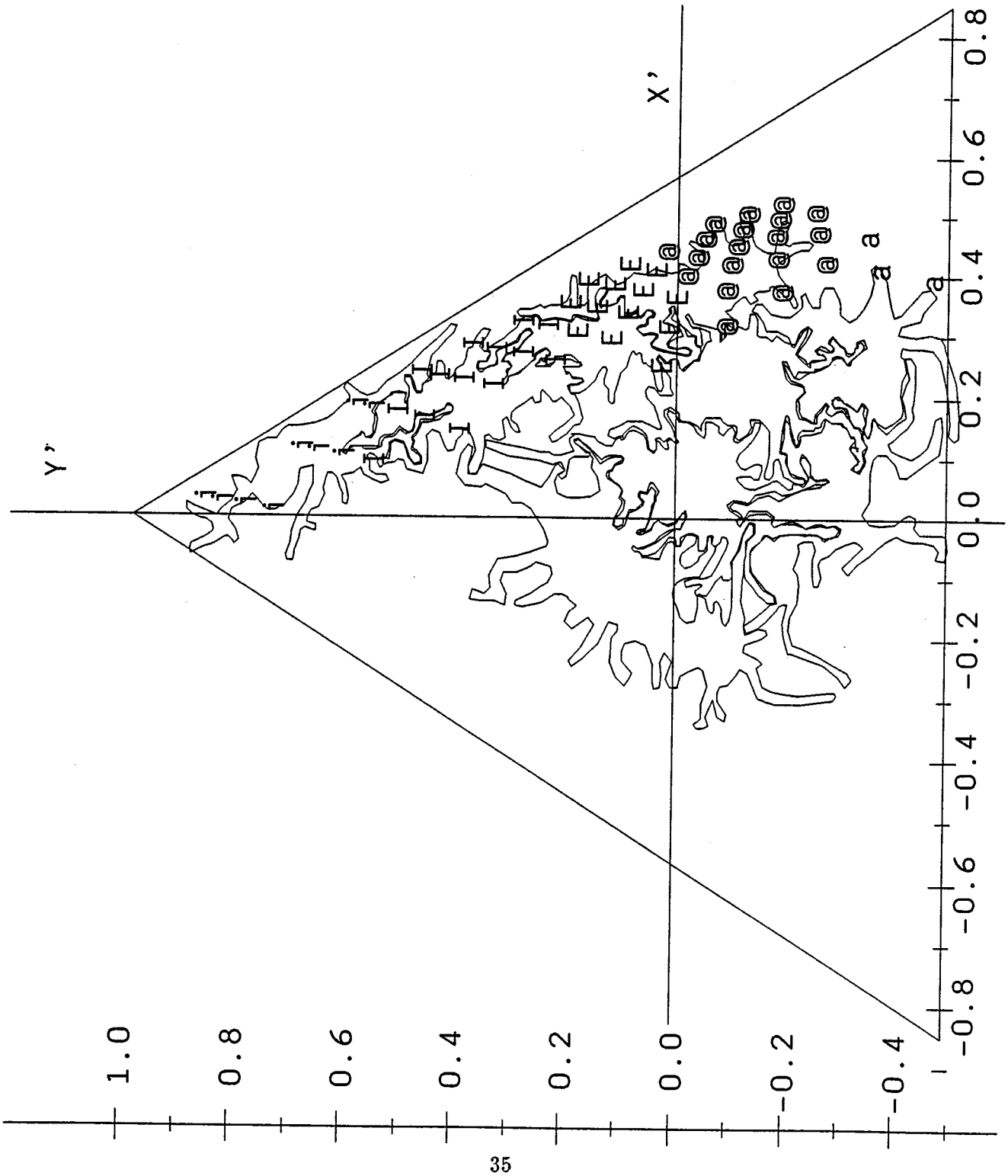
Only a few studies have utilized a relatively wide range of possible F1-F2 combinations (Holmes, 1986; Scholes, 1967; Ainsworth and Millar, 1971; Millar and Ainsworth, 1972;

Nearey, 1977; R.L. Miller, 1953). Of these, only R.L. Miller (1953) and J. Holmes (1986) included a variable F3 for some stimuli. Unfortunately, subjects in the Holmes (1986) study were instructed to identify only tokens judged to be exactly phonetically correct and thus other additional information relevant to the vowel categories and their boundaries was lost. However, in the R.L. Miller (1953) study, several hundred two- and three-formant vowel tokens were synthesized with a man's (144 Hz) and child's (288 Hz) fundamental frequency and presented to subjects for identification. Among the findings were that (1) general areas assigned to vowels as plotted in an $F1 - F2$ space remained relatively fixed for sounds of different fundamental frequencies; (2) the addition of a third formant added considerably to the unanimity of responses, particularly for front vowels; and (3) the presence or absence of sounds from certain regions may influence a subject's response to sounds of other regions. Visual inspection of the $F1 - F2$ plots from this study suggests that (1) boundaries between vowels may be abutting and irregularly shaped, and (2) these boundaries may be quite narrow. The nature of these boundaries however, was not the focus of this study. Results from this study of identifications for stimuli utilizing a third formant have been plotted as points in the auditory-perceptual space (Figure 2-1) and approximately 85% of the identifications are found to be either in agreement with the current *PTZ* estimates or lie in adjacent unclaimed space.

Given these motivations, several categories of questions may be addressed. The first category pertains to the perceptual responses themselves. Given that all tokens can be identified as vowels of American English, what acoustic variables of the stimuli mediate these responses? How well do subjects agree on identifications and confidence ratings and how reliable are these agreements? Are identification agreements and confidence ratings correlated or do they reflect different kinds of information about the stimuli? What role do individual differences play in the results?

A second category of questions is directed toward the concept of vowel zones in the *APS*. Can zones for the vowels of American English be constructed on the basis of the results of such an experiment? If so, what are the locations and shapes of the zones when all synthetically realizable space in the *APS* is considered, and how do they compare to the *PTZs* based on natural speech? What is the nature of the boundaries between the zones,

Figure 2-1: Subjects' identifications for synthetic vowels from R.L. Miller (1953) plotted in APS $x'y'$ coordinates along with current target zone estimates from Figure 1-3.



that is, how well defined are they, and how do they vary between listeners? Can a measure of vowel "saliency" be derived from the responses, and if so, how is saliency distributed within and between the zones?

The last category addresses the concept of target zones in the *APS* as a classification scheme for vowels. How do target zones based on the results of this experiment compare to other vowel classification schemes in their ability to correctly classify synthetic and natural vowels ?

2.2 Methods

2.2.1 Stimuli

Isolated vowel sounds representing points in the *APS* were synthesized with 16-bit resolution at a 10 kHz sampling rate using the cascade portion of a Klatt digital formant synthesizer (Klatt, 1980), implemented on a DEC MicroVax II computer. These points are equi-distant .05 log unit (whole tone) steps in the $x'y'$-plane and arbitrarily originate from x'= 0.0, y'= 0.0, such that all coordinate values in both dimensions are evenly divisible by 0.05. Figure 2-2 shows the points for one z' slab ($z' = 0.70$). The points span seven slabs in the $z'y'$-plane (Figure 2-3), ranging from $z' = 0.50$ to 0.80 in 0.05 log unit steps. This range is based on the z' values for data points in the CID natural speech database[1] and the Peterson and Barney (1952) study where the predominant z' range is 0.65 to 0.75 for non-retroflex vowels. However, to include consideration of the retroflex /ER/, the range of z' must extend back to at least $z' = 0.53$ in accordance with the locations of natural data points. Thus the range of $z' = 0.50$ to 0.80 ensures that the range of vowel sounds most likely to be produced in natural speech will be included among the synthesized stimuli. However, the three planes where $z' = 0.65$, 0.70 and 0.75 will be referred to as the *primary* planes and the planes where $z' = 0.50$, 0.55, 0.60, and 0.80 the *secondary* planes.

The 0.05 log unit resolution is thought to be poorer than that required to adequately determine *PTZ* boundaries, but sufficient to provide an indication of boundary areas between *PTZs*, as well as the general locations of the *PTZs* with a reasonable number of stimuli.

[1]For a detailed description of this database, see Miller, 1989, Appendix B

Figure 2-2: Location in APS $x'y'$ coordinates of synthesizable tokens for one z' plane ($z' = 0.700$).

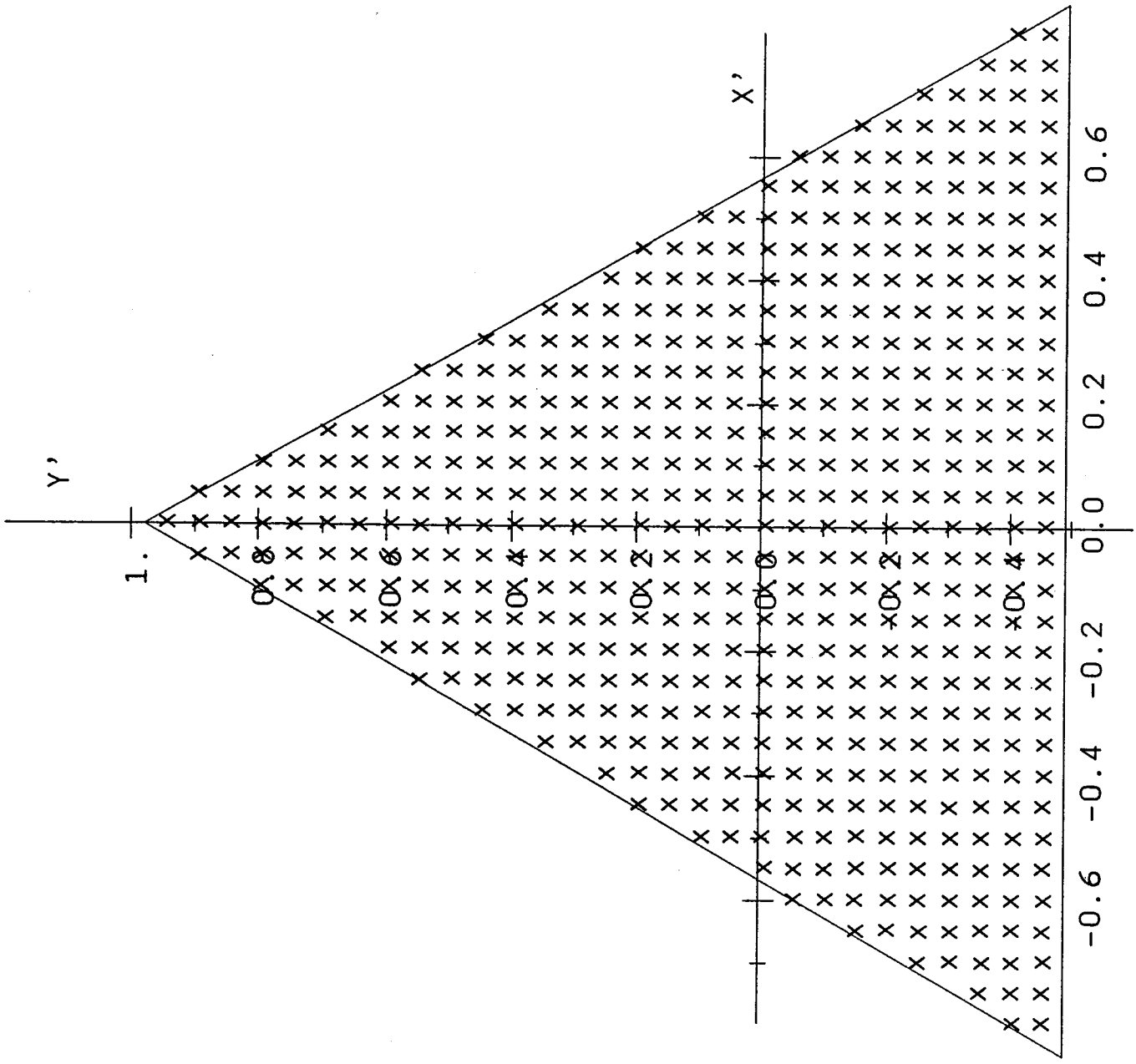
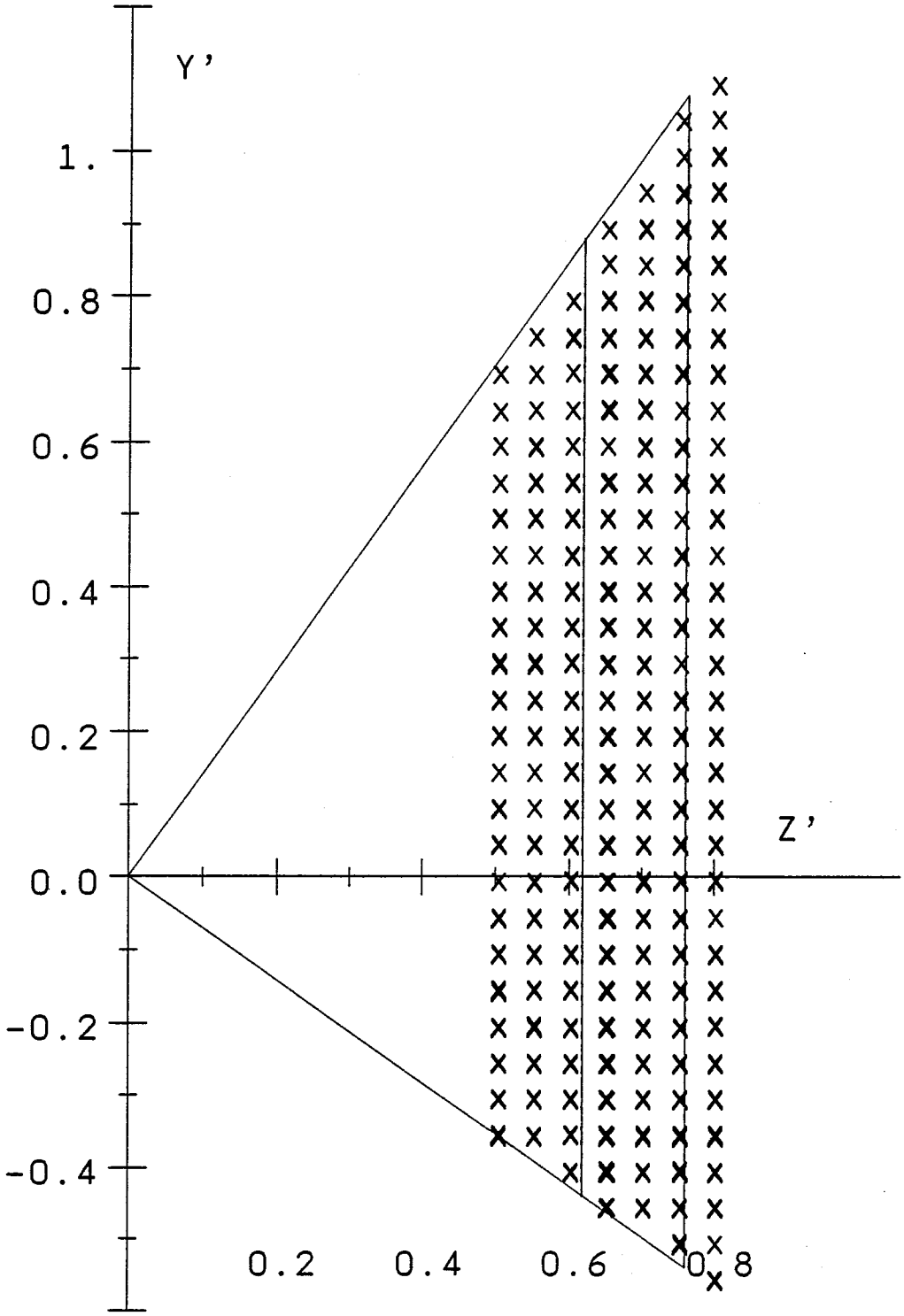


Figure 2-3: Location in APS $y'z'$ coordinates of z' planes utilized for Experiment I.



Valid points were initially considered to be those where $SR \leq F1 \leq F2 \leq F3$. This criterion yields 3151 possible points for synthesis. However, results from pilot studies indicated that a number of these synthetic tokens, particularly those near the outer borders of the vowel space, sound very unnatural or "unspeechlike." To eliminate these tokens and to reduce the total possible tokens to a more manageable number, a second set of criteria was employed. Rules based on the CID natural-speech-data corpus for specifying the possible frequency-band ranges of $F1, F2$, and $F3$ relative to SR have been developed for possible use in an automatic formant-picking procedure. These formant-band range (FBR) estimates indicate the minimum and maximum allowable values of $log(Fn/SR)$, where $n = 1$, 2 or 3. The current estimates of these values[2] are,

$$\begin{aligned} 0.09 &\leq log(F1/SR) \leq 0.85 \\ 0.56 &\leq log(F2/SR) \leq 1.20 \\ 0.86 &\leq log(F3/SR) \leq 1.40. \end{aligned} \tag{2.1}$$

By using these rules as additional criteria, the total number of acceptable tokens was reduced by 45% to 1725. Figure 2-4 shows the acceptable points in the same z' slab as in Figure 2-2 plotted over the most current estimates for PTZ boundaries from Figure 1-3 in Chapter 1. Note that virtually all the space occupied by the $PTZs$ is acceptable for synthesis.

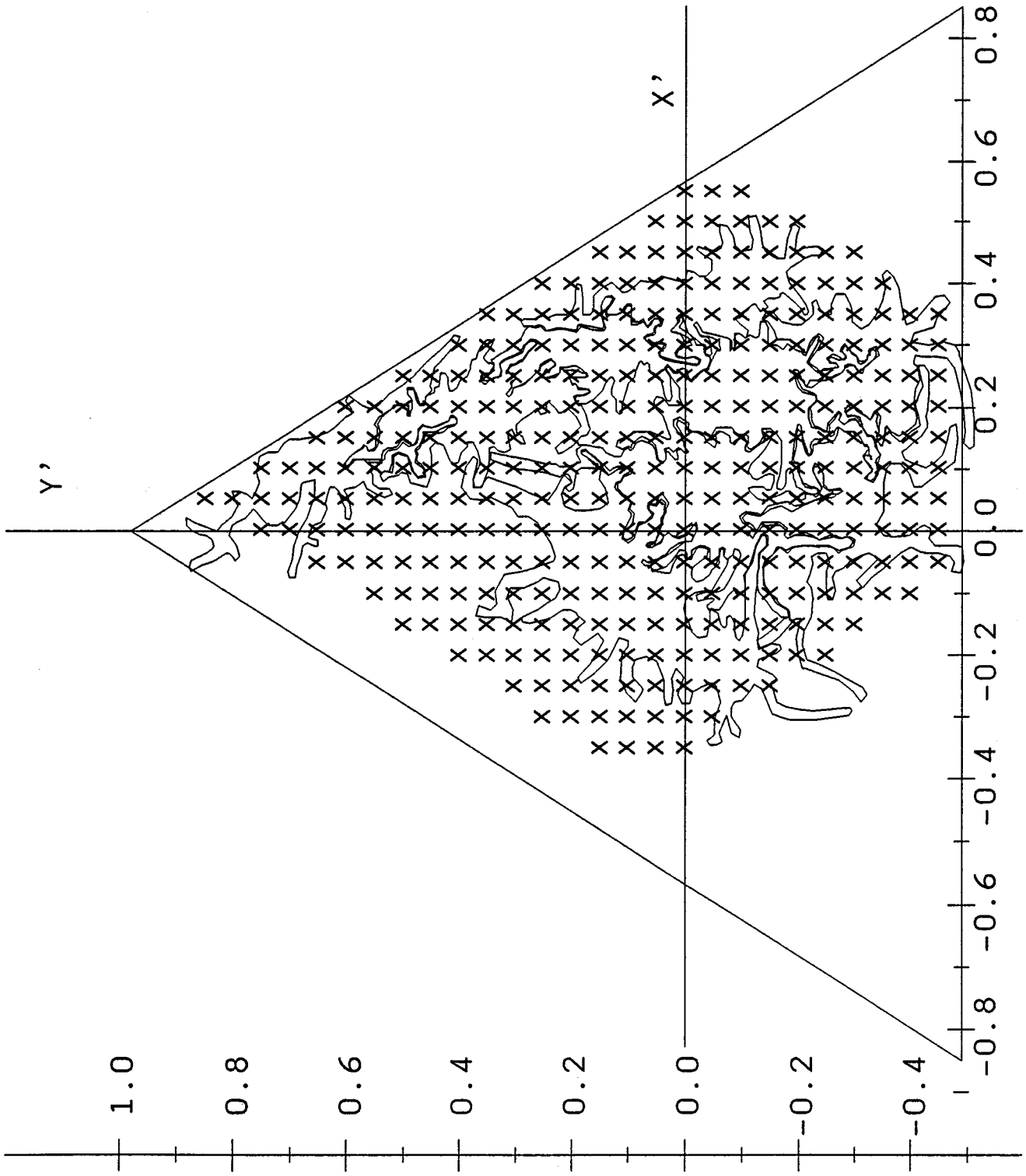
Frequency values for $F1, F2$, and $F3$ were calculated from each set of x', y', z' coordinates by means of a matrix equation. The value of the sensory reference (SR) was held constant at 155 for all tokens, yielding a fundamental frequency ($F0$) of 132 Hz and FBR ranges of

$$\begin{aligned} 190 < F1 < 1097 \\ 563 < F2 < 2456 \\ 1122 < F3 < 3890. \end{aligned} \tag{2.2}$$

$F4$ is arbitrarily set to 4000 Hz, somewhat higher than would be typically found for a male talker, to accomodate the use of higher-than-normal values of $F3$. All tokens were 400 ms in duration[3]. The amplitude contour of each token was ramped by logarithmically

[2]The minimum value of the FBR for $F1$ is normally set to 0.00 to provide for instances where $F1$ may drop to the value of SR as in voice bars. However, for vowels, $F1$ does not fall below 190 Hz for any data found in the CID data corpus. Thus, the minimum value of the FBR for $F1$ has been set to 0.09 ($log(190/155)$) for the purposes of this experiment.

Figure 2-4: Locations in $x'y'$ of tokens in one z' plane ($z' = 0.700$) acceptable for synthesis after applying formant-range limiting criteria plotted with the most recent target zone estimations from Figure 1-3.

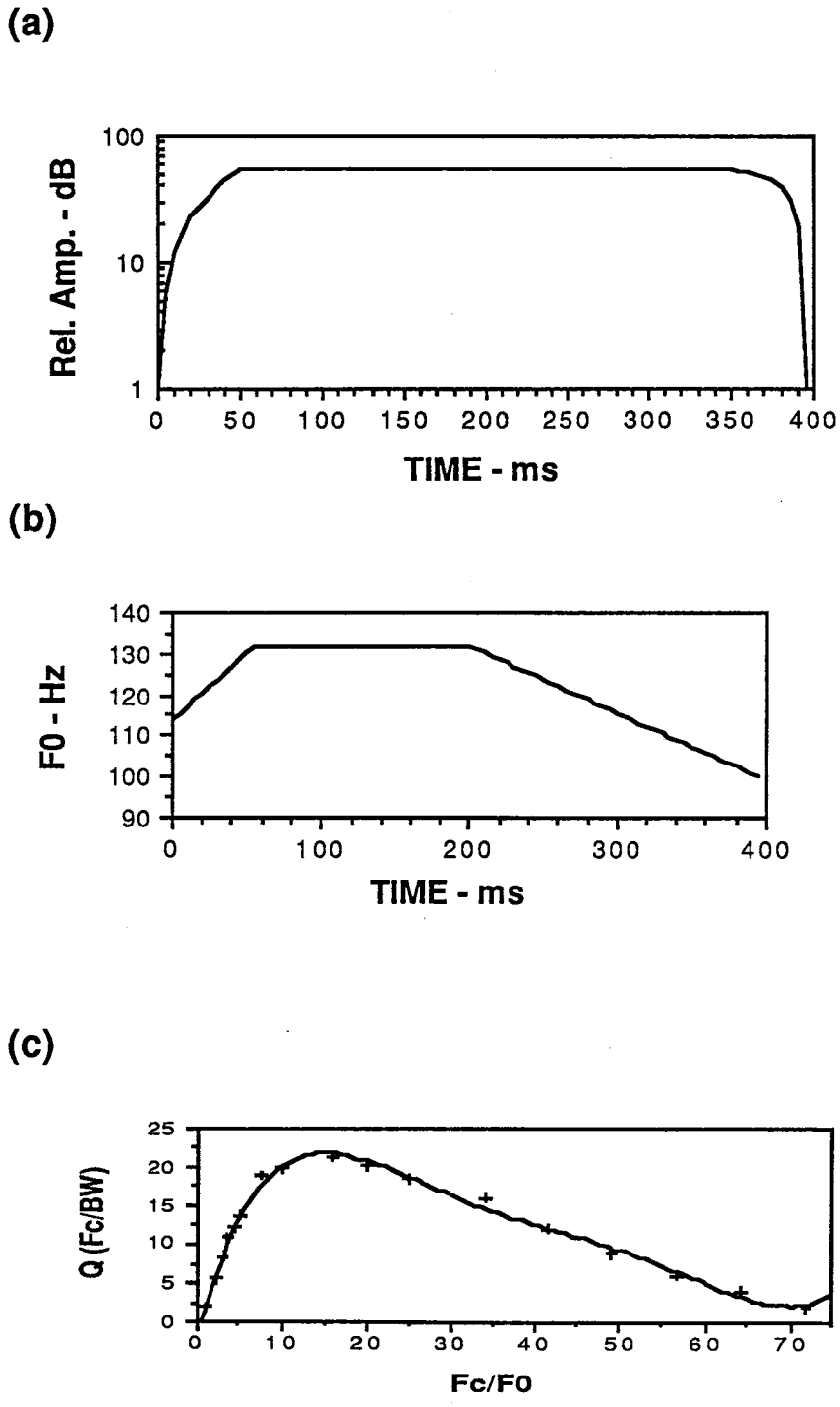


interpolating from 1 to 55 dB over the first 50 ms, held at a constant 55 dB for the next 300 ms, and again ramped by log interpolation from 55 to 1 dB over the last 50 ms (Figure 2-5a). The $F0$ contour of each token followed approximately a scaled version of the parameters used by Burdick and Miller (1975), linearly interpolating from 114 to 132 Hz over the first 50 ms, maintaining a steady-state 132 Hz over the next 150 ms, and again linearly interpolating from 132 to 100 Hz over the last 200 ms (Figure 2-5b). Amplitude normalization across tokens was achieved by means of scaling all tokens to equal peak amplitude.

Values for the first three formant bandwidths ($B1, B2, B3$) were generated by means of an equation based on data from Miller (1980). In this study, bandwidth (BW) estimates for males from Dunn (1961) and Fujimura and Lindqvist (1971) were averaged and plotted as Q (Fc/BW) over the ratio of the formant center frequency (Fc) to the fundamental frequency ($Fc/F0$). These data were then fit with a curve by means of a 5th-order polynomial regression (See Figure 2-5c). The coefficients from this regression are used to calculate the formant bandwidths. The bandwidth values generated by this equation agree reasonably well with other published values and predictive formulas (House and Stevens, 1958; Fant, 1972). All other synthesis parameter specifications are listed in Appendix B. Examples of spectral envelopes from the stimuli are shown in Appendix C.

[3]The lax vowels of English [IH, EH, AE, AH, UH], with the exception of [AE] are often considered the short vowels and their duration in natural speech averages about 80% of the vowels considered intrinsically long in English, [IY, AA, AO, UW], even when spoken in isolation (Strange, Edman, and Jenkins, 1979). The differences in vowel duration have long been considered a "secondary cue" to vowel perception (Peterson and Lehiste, 1960). However, past studies addressing this issue have reported conflicting results. Ainsworth (1972) demonstrated that duration can effect the identification of synthetic vowels. This result must be viewed in light of the fact that two-formant vowels were utilized in this study which may have sounded less natural and perceptually less salient than natural speech, therefore potentially increasing the weight given to durational cues as a response criterion. Additionally, Ainsworth (1972) suggests that the importance of duration as a cue may be less important to isolated vowels than vowels in a linguistic context. Pisoni (1971), Stevens, et al. (1969), and Repp, et al., (1979) all found that the [IH] category identification responses to be less stable than the [IY] or [EH] responses and speculated that this instability could be due to the relatively long stimulus durations compared to natural speech. However, the current experiment is not as concerned with response stability within vowel categories as it is with the stability of boundaries between categories. The three studies previously mentioned all found relatively sharp boundaries between categories at approximately the same place in the vowel stimuli continuum. Additionally, Pisoni(1973) demonstrated that this same boundary (for IY-IH) sharpness and location remains stable with synthetic vowels of both long (300 ms) and short (50 ms) durations.

Figure 2-5: (a) Overall amplitude contour used for token synthesis; (b) Fundamental frequency ($F0$) contour used for token synthesis; (c) Q as a function of the ratio of formant center frequency (Fc) over fundamental frequency ($F0$) used for formant bandwidth calculation in token synthesis (See text).



2.2.2 Procedure

Identification and confidence-rating tasks were employed for the mapping of the *APS* vowel slab. The tokens were randomized, low-pass filtered at 5 kHz, and presented via a digital-to-analog converter (MicroTechnology Unlimited DigiSound-16) directly from the computer. Subjects listened binaurally by headphone (AKG K-141) in a sound-attenuated room at a comfortable listening level ($\approx$ 55-60 dBA-Slow SPL). Subjects were asked to identify each token as one of the following vowels, [IY, IH, EH, EY, AE, AA, AH, AO, OW, UH, UW, ER]. Although [EY] and [OW] are generally considered diphthongal nuclei and not pure vowels in American English, they are included here to allow more thorough comparison with previous mapping experiments (Miller, 1953) and because they may be considered phonetically as single target speech sounds in American English (Lehiste and Peterson, 1961). Currently no target zones exist for [EY] and [OW], so estimations of these zones will be required.

In an earlier pilot study, subjects were asked to rate each token for its "goodness" on a 9-point scale. However, these subjects indicated that the nine-point goodness-rating scale was too broad to establish and monitor, suggesting that a rating scale using fewer categories should be considered. Furthermore, it was determined that the "goodness" rating reflected too many different variables, so the rating criteria were reduced to reflect only the certainty of the response. Instead of a goodness rating, subjects were asked to rate each token identification response for "confidence" on a 5-point scale. Subjects were instructed that the confidence rating should reflect how certain they were as to their identification of the token. A rating of 1 should reflect that the subject was very unsure as to the category they selected, a rating of 2, 3, or 4 should reflect moderate levels of certainty, and a rating of 5 should reflect that the subject was very confident of their response.

The 1725 tokens were randomly divided into 17 blocks with 16 blocks of 100 tokens each and 1 block containing 125 tokens. The presentation order of the blocks was randomized per subject as well. Sessions were under computer control and structured in the following manner. A row of 12 boxes appeared on the terminal screen. Within each box was a two-character arpabet symbol for one of the categories and a [hVd] word containing that vowel sound. A message flashed on the screen that the token would be presented in two seconds. After the token was presented, the subject was prompted for a response. The identification

responses were made by pressing keys labeled with the same two-character symbols. After the identification response was entered, the subject was prompted for a confidence rating response. A row of boxes containing the numbers 1 through 5 with box 1 labeled "very unsure," box 3 labeled "mediocre," and box 5 labeled "very sure" appeared and the subject selected their response by numerical key entry. This concluded a single token presentation. Subjects had the option of repeating the token being judged a number of times if necessary at any point in the process prior to giving the rating response. The number of times a token was repeated was also recorded. Subjects kept a written account of errors they had made in key entry which the experimenter later corrected. Each subject evaluated all 1725 tokens twice.

2.2.3 Subjects

Subjects, five male and five female, were recruited from the student body of Washington University and the nearby St. Louis area. Subjects' ages ranged from 17 to 25 years. All subjects were native speakers of American English with no known history of either speech or hearing impairment. All subjects were naive in terms of formal phonetic training. Six of the subjects were born and raised in the Midwest and the remaining four had resided in St. Louis a minimum of two years. While some dialectal differences between subjects were anticipated, all subjects were screened to ensure that he or she normally used and perceived in their everyday speech all of the vowel sounds to be used as response categories in the experiment. Two subjects, one male and one female, were unable to complete the experiment, thus the results presented here will reflect the data collected from eight subjects.

Training

Subjects were trained in the experimental protocol in two stages. In the first stage, subjects trained on 51 tokens consisting of: 1) synthetic vowels constructed using the male average formant frequencies values from Peterson and Barney (1952); 2) exemplar tokens selected from the test stimuli; and 3) vowels spoken and recorded by the experimenter imitating the test stimuli. Subjects identified ten randomizations of these tokens or achieved a consistent identification rate of at least 96% correct, whichever came first.

In the second stage of training, subjects identified and provided confidence ratings for the 304 test tokens from the $z' = 0.70$ plane later in the main experiment. This plane lies midway along the z' dimension in the vowel slab and vowel tokens synthesized from this plane generally evoke salient perceptions of eleven of the twelve categories[4] used for identification. The 304 tokens were randomized and divided into three blocks. Subjects' identifications were analyzed to ensure that all eleven categories were used and that there was a general consistency of grouping for like identifications . The results of this informal analysis were discussed with the subject. One subject (1M) did not receive this portion of the training. This concluded the second stage of training.

2.3 Results

2.3.1 General observations

A number of observations were noted from conversations with subjects following the experiment which are of interest here. Subjects noted that a number of tokens did not sound like vowels of English, but rather, like the front rounded vowel /ü/ used in German, and that their perception of these tokens was ambiguous between /IY/ and /UW/. Examination of the data and informal listening by trained phoneticians indicated that these tokens fell in the area at the border of /IY/ and /UW/. Thus, a region along this border most probably should be classified as "not a vowel of English."

Although the diphthong /EY/ can be produced as a monophthong in American English (Pike, 1947; Lehiste and Peterson, 1961) and occurs as a pure vowel in many languages, several subjects noted that they had difficulty in distinguishing /EY/ from /IH/ and /EH/. This difficulty might have been reduced had more rigorous training in identifying this category been employed. Monophthongal versions of /OW/ did not seem to present such a problem.

Subjects also noted that a number of tokens sounded like something between a purely monophthongal vowel and /ER/, making these tokens difficult to classify. Informal listening by the experimenter indicated that tokens falling near the boundaries for /ER/ tended to

[4]The vowel /ER/ does not occur in this z' plane.

sound very rhotacized, or "r-colored", presumably creating the classification difficulty.

Before the analysis of the data in perceptual terms, it is desirable to know how subjects utilized the identification and confidence rating categories. Specifically, do the subjects represent a homogeneous group in terms of the identification and confidence rating responses? Are subjects' responses reliable in terms of consistency and repeatability?

2.3.2 Identifications

The frequency of responses for each identification category are shown in Table 2.1 for each subject's response sets. The subject response sets are designated by first the subject number, followed by M or F (for male or female), and then the set number (1st or 2nd replication). These designations will be referred to throughout the results section. The range of response frequencies is great, varying from as few responses as 9 (3M1-/EY/) to as many as 513 (1F1-/UW/). Category /EY/ had the fewest average number of responses overall and /UW/ the greatest with the remainder of response categories falling roughly into two groups, /IH,EH,AE,ER/ and /IY,AA,AH,AO,OW,UH/.

The average percentage of agreement between identifications collected from each subject response set paired with all other subject response sets for all 1725 tokens across all confidence ratings are shown in Table 2.2. The average percentage of agreement across all subjects was 63.8% and drops to 62.9% when the agreements of subjects with their own replications are not included. These averages and their standard deviations for agreement with others are plotted in Figure 2-6 for all subject response sets. The average agreement across subjects for the first response sets was 61.6% and for the second response sets increased to 64.5%.

To test differences in agreement, the statistic kappa (κ) (Cohen, 1960) was utilized. This statistic provides a coefficient of agreement between two raters for nominal scales and includes consideration of chance agreements. The statistic assumes that all disagreements may be considered equally serious. A coefficient of 1.00 reflects perfect agreement between raters, while a coefficient of zero reflects total independence between the two raters. When N is large (i.e. > 100),the sampling distribution of κ approximates normality. Thus differences between two values of κ may be tested for significance by evaluating the normal curve

Table 2.1: Frequency of ID responses by subject response set.

Set	IY	IH	EH	EY	AE	AA	AH	AO	OW	UH	UW	ER
1M1	53	63	63	22	88	174	130	170	146	269	397	150
1M2	69	63	89	41	100	167	129	170	151	256	394	96
3M1	135	187	73	9	68	166	176	138	130	223	374	46
3M2	136	164	69	14	56	132	161	144	153	232	385	79
4M1	269	122	77	17	102	142	104	152	171	108	245	216
4M2	264	143	95	18	102	118	113	146	183	128	271	144
5M1	239	106	52	149	158	186	120	122	86	120	306	81
5M2	269	125	47	119	141	194	116	117	129	87	314	67
1F1	82	25	40	37	83	123	130	256	120	225	513	91
1F2	178	116	65	22	77	141	120	192	137	196	398	83
2F1	217	73	86	64	94	133	127	144	171	156	377	83
2F2	289	59	106	83	81	125	116	135	199	160	331	41
3F1	95	66	99	16	70	113	164	143	161	312	375	111
3F2	90	61	102	21	58	92	155	175	166	253	454	98
5F1	174	14	118	89	85	133	197	160	180	56	448	71
5F2	203	20	117	79	108	152	173	132	198	85	413	45
$\bar{x}$	173	88	81	50	92	143	139	156	155	179	375	94

deviate.

Kappas and standard errors of κ were calculated for each subject's first response set paired with all other first response sets and similarly for subjects' second response sets. The average value of κ for the first response sets was .570 and for the second response sets was .602. The difference between these values was not found to be significant ($z = 1.78$), suggesting that while subjects' agreements improved in the second set, they were not significantly different from the first set agreement performance.

To consider test-retest reliability, the percentage of agreement between each subject's first and second response set was first calculated. The average of these agreements was 75.31%, approximately 12% greater than the overall average of agreements between each subject and all others. Kappas and standard errors of κ were calculated for each subject's response set paired with all other response sets except the subject's replication. The average across all subjects was $\kappa = .585$. The same statistics were calculated for each subject and their replication, yielding an average $\kappa = .721$. The difference between these two values of κ was tested for significance by evaluating the normal curve deviate. The difference was highly significant ($z = 7.76, p < .001$). This result suggests that test-retest reliability within subjects is quite high when compared to average agreement between subjects.

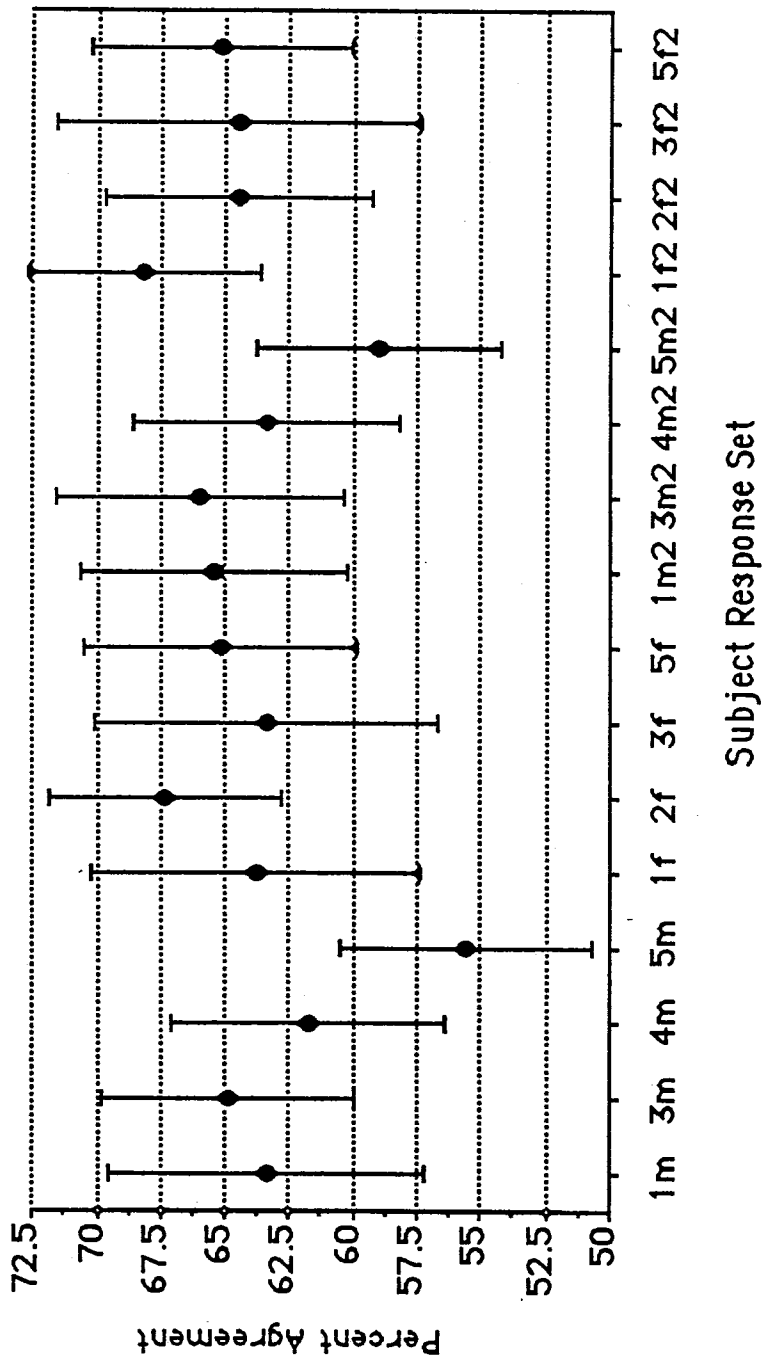
2.3.3 Ratings

The frequencies of rating responses by rating category are shown in Table 2.3 for each subject response set. It is readily apparent that the distribution of rating responses is negatively skewed with the highest frequency of ratings occurring in the fourth category representing a confidence level of greater than mediocre but less than very sure. It does appear that subjects were generally more sure of their responses than less sure. The variations in the frequency of rating responses within subjects is smaller than the variations between subjects. This suggests that each subject establishes his/her own criteria for ratings judgements, and maintains the same criteria run to run. This notion is consistent with data reported in a study investigating the use of confidence ratings for identifications of spondee words presented in noise by F.R. Clarke (1960). As was previously reported by Pollack and Decker (1958), Clarke noted that listeners tended to underestimate their ratings of high

Table 2.2: Percentages of identification agreement by subject response set.

Run	1M	1M2	3M	3M2	4M	4M2	5M	5M2	1F	1F2	2F	2F2	3F	3F2	5F	5F2
1M	—															
1M2	75.4	—														
3M	65.1	67.7	—													
3M2	66.1	67.9	76.2	—												
4M	58.2	59.4	60.6	61.2	—											
4M2	59.0	60.9	64.2	65.9	76.4	—										
5M	52.2	56.0	55.8	54.2	55.8	55.8	—									
5M2	55.4	58.7	59.3	57.9	58.9	59.8	70.4	—								
1F	70.4	70.7	64.3	65.7	57.0	58.5	51.9	54.0	—							
1F2	70.1	71.5	72.0	73.8	65.0	67.0	57.5	60.8	71.9	—						
2F	63.8	66.3	67.2	69.9	67.7	69.0	57.9	61.7	65.2	70.5	—					
2F2	58.2	63.1	64.5	66.6	66.0	67.0	56.5	60.8	59.4	67.8	76.8	—				
3F	65.8	66.1	63.1	67.0	59.5	61.8	50.0	51.7	66.2	66.4	65.2	60.2	—			
3F2	66.5	68.0	65.6	68.9	59.1	61.0	48.9	51.5	71.8	68.6	66. 6	62.4	76.2	—		
5F	62.6	64.8	63.5	64.5	61.4	61.2	55.5	61.2	65.6	67.6	70.9	68.8	63.9	66.8	—	
5F2	61.6	64.7	64.0	63.9	61.7	63.0	57.3	63.4	64.1	67.0	71. 0	68.8	62.7	65.0	79.2	—
$\bar{x}$	63.4	65.4	64.9	66.0	61.8	63.4	55.6	59.0	63.8	67.8	67. 3	64.4	63.0	64.4	65.2	65.2

Figure 2-6: Mean agreement between each subject response set and all other response sets on identification of the 1725 synthetic vowels. Error bars indicate ±1 standard deviation.



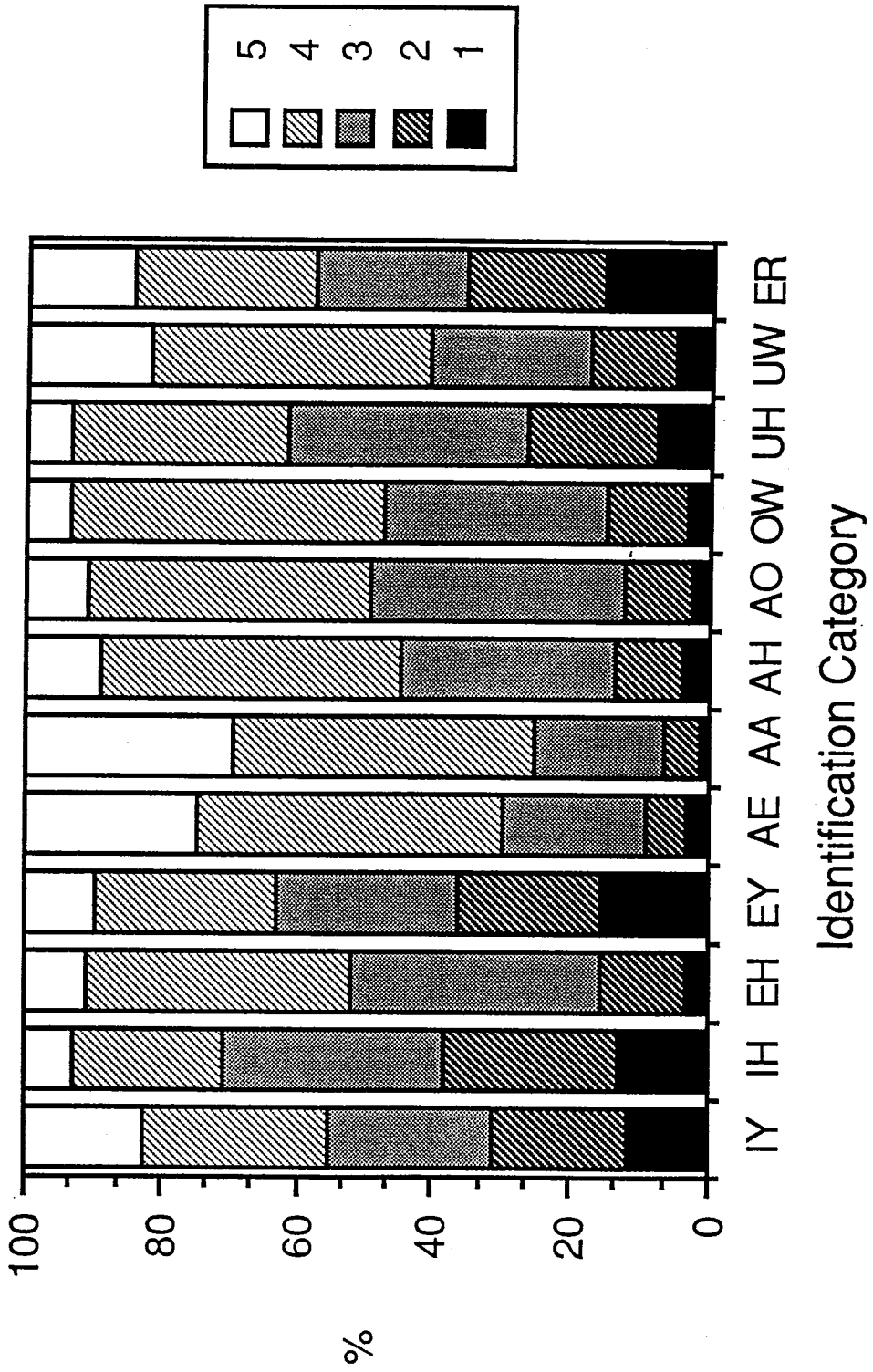
intelligibility items (and overestimate their ratings of low intelligibility items). A pattern of this nature could account for the high frequency of ratings in category 4 of the present data if subjects are, in fact, underestimating their confidences at identifying tokens of high salience.

Table 2.3: Frequency of rating responses by subject response set.

Set	1	2	3	4	5
1M	63	207	444	695	316
1M2	47	164	370	647	497
3M	87	369	636	548	85
3M2	73	349	680	548	75
4M	214	257	371	692	191
4M2	193	201	407	735	189
5M	236	418	373	360	338
5M2	171	325	408	380	441
1F	34	132	488	618	453
1F2	9	130	424	658	504
2F	54	183	441	725	322
2F2	57	193	429	843	203
3F	220	273	679	549	4
3F2	83	213	509	909	11
5F	120	197	552	749	107
5F2	90	156	469	757	253
Total	1751	3767	7680	10413	3989
$\overline{x}$	109.4	235.4	480.0	650.8	249.3

Figure 2-7 shows the frequencies of rating responses for each identification category expressed as percentages of all ratings for each category. The figure indicates that the majority of tokens across all categories received a rating of 3 or 4, as was noted above,

Figure 2-7: Percentage of subjects' confidence rating responses by individual identification category.



but also, that tokens classified as the "point" vowels /IY,AE,AA,UW/ and the retroflex /ER / received greater than 15% of their ratings in the "very sure" (5) rating category, suggesting that a greater percentage of tokens classified as these vowels were very salient as opposed to tokens classified as more centralized vowels. Additionally, tokens classified as /IY,IH,EY,UH,ER/ received greater than 25% of their ratings in rating categories 1 and 2. These higher percentages of low confidence ratings most probably reflect observations noted in Section 2.3.1, that is, for the /IY/ category, a number of tokens probably most representative of a rounded front vowel /ü/ not used in American English may have been classified as /IY/. Likewise, uncertainty in classifying /EY/ may reflect that many subjects expressed difficulty in differentiating /EY/ from neighboring vowels. Classifying tokens in the /ER/ region may present a higher percentage of uncertainty in that tokens synthesized near the /ER/ boundary take on a high degree of retroflexion and are very "r-colored", creating ambiguities. The lax vowel categories, in particular /IH/, have often appeared less stable in terms of perceptual identification (See footnote 2, Section 2.2.1) than are other vowel categories of English, suggesting that a higher percentage of uncertainty may be associated with them.

The percentage agreement on confidence ratings between each subject response set and every other response set is shown in Table 2.4. The overall average agreement is 30.2%, considerably less than their agreement on token identifications. This overall agreement drops to 29.6% when the agreements of subjects with their own replications are not included. The average value of κ for these agreements is quite small, .049, suggesting that a large portion of agreements may now be random chance. Subjects' average percentage of agreement was uniform across the first (29.2%) and second (30.0%) response sets. Once again, the agreement within subjects (38.6%) was significantly higher than the overall average based on a comparison of the average κ statistic for each condition ($z = 4.38, p < .001$).

In summary, the data from confidence ratings suggests that subjects establish individual criteria for basing their confidence judgements, and then use these criteria consistently. These individual differences are reflected in the lower agreements between subjects found for confidence ratings as compared to identification agreements. The majority of ratings fall in category 4, reflecting a confidence judgement which is greater than mediocre but

Table 2.4: Percentages of agreement on confidence ratings by subject response set.

Run	1M	1M2	3M	3M2	4M	4M2	5M	5M2	1F	1F2	2F	2F2	3F	3F2	5F	5F2
1M	—															
1M2	38.2	—														
3M	32.3	27.5	—													
3M2	32.5	25.3	44.0	—												
4M	31.4	27.0	30.1	28.8	—											
4M2	31.7	27.8	30.3	31.2	36.1	—										
5M	28.3	26.7	29.0	25.8	25.2	25.2	—									
5M2	27.7	26.4	25.4	26.1	25.0	23.6	36.3	—								
1F	32.2	34.7	27.1	27.3	26.5	28.1	23.2	21.9	—							
1F2	35.9	36.8	28.3	27.0	28.8	31.1	25.7	26.6	38.1	—						
2F	36.4	37.8	31.9	31.2	32.1	31.5	25.0	26.1	35.5	34.7	—					
2F2	35.4	33.9	35.0	32.5	33.2	35.2	26.2	26.4	33.0	35.5	43.2	—				
3F	28.1	23.5	32.1	33.7	28.6	29.8	21.9	22.4	25.2	23.9	27.4	29.8	—			
3F2	33.0	28.4	35.4	35.7	32.3	33.9	21.8	20.9	30.7	29.5	32.9	35.8	37.4	—		
5F	32.1	29.4	30.1	31.7	29.4	34.5	22.3	20.2	29.6	28.3	32.8	35.1	30.6	36.6	—	
5F2	32.3	28.9	28.8	29.2	29.4	31.7	20.9	22.9	32.5	30.6	34.5	35.7	29.2	33.6	35.2	—
$\bar{x}$	33.8	30.1	31.2	30.8	29.6	30.8	25.6	25.2	29.7	30.7	32. 9	33.7	28.2	31.9	30.5	30.4

less than very sure. Results of past studies, however, suggest that subjects may tend to underestimate their confidence when they are correct. The identifications assigned to vowel categories representing the extreme points of articulation, (/IY,AE,AA,UW,ER/) tended to receive higher confidence ratings than did vowel tokens in categories representing less extreme articulatory points.

2.3.4 Synthetic speech-based (*SSB*) target zones

Consideration is next given for whether target zones similar to those illustrated in Figures 1-8 and 1-9 from Chapter 1 can be constructed in the *APS* based on the identifications for the 1725 synthetic tokens. A single vowel category classification for each stimulus token is required for constructing such zones. The vowel category representing the plurality of identification responses for each token was determined for this purpose. That is, for each point in the *APS* representing one of the tokens, the vowel identification category receiving the greatest number of the 16 (8 subjects x 2 replications) possible responses for that token was taken as the category label for that point.

Boundary lines were drawn enclosing points of like plurality identification in each z' plane. Each point was considered to represent the space occupied by a 0.05 log unit cube with the point at its center. Adjacent cubes of unlike categories were separated by boundary lines drawn equidistantly between the two cubes such that each line was .024 log units away from its associated cube center in the $x'y'$-plane. This technique successfully enclosed 1674 of the token points into adjacent, non-overlapping zones along each of the five z' planes and are shown in Figures 2-8a-g.

Of the 51 token points remaining, 49 represented ties between the greatest number of subjects' responses for two or more categories such that no clear plurality could be established. Of the tied token points, 3 represented ties between three identification categories and the remaining 46, ties between two categories. All but one of the tied token points fell along boundary areas between the adjacent zones which were representative of the identification categories of the ties. These points were included in the zone construction by drawing boundary lines diagonally to either side of the point in the $x'y'$-plane or otherwise dividing the point's cube such that all tied response categories associated with that cube

Figure 2-8: (a) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.80$ plane.

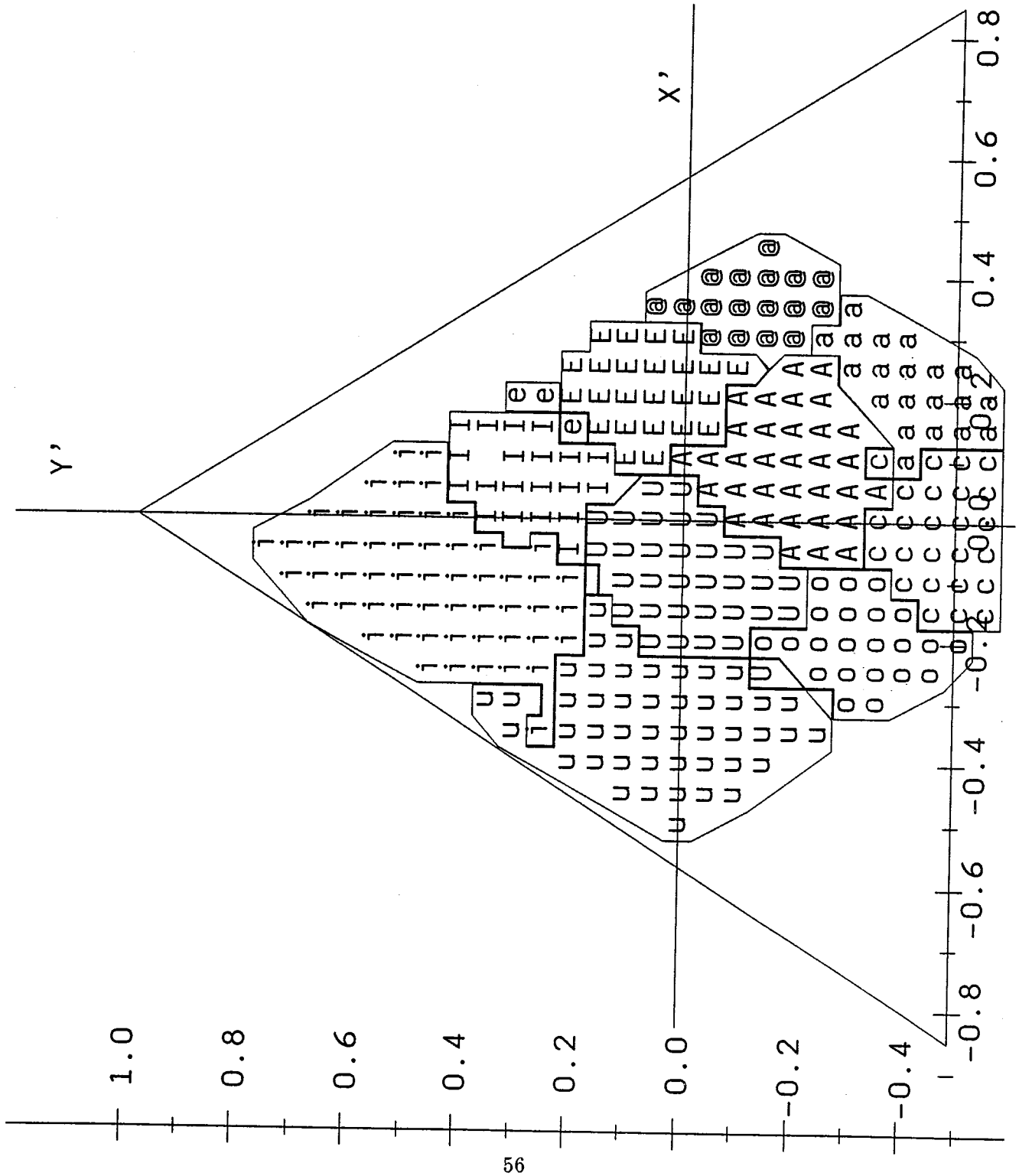


Figure 2-8: (b) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.75$ plane.

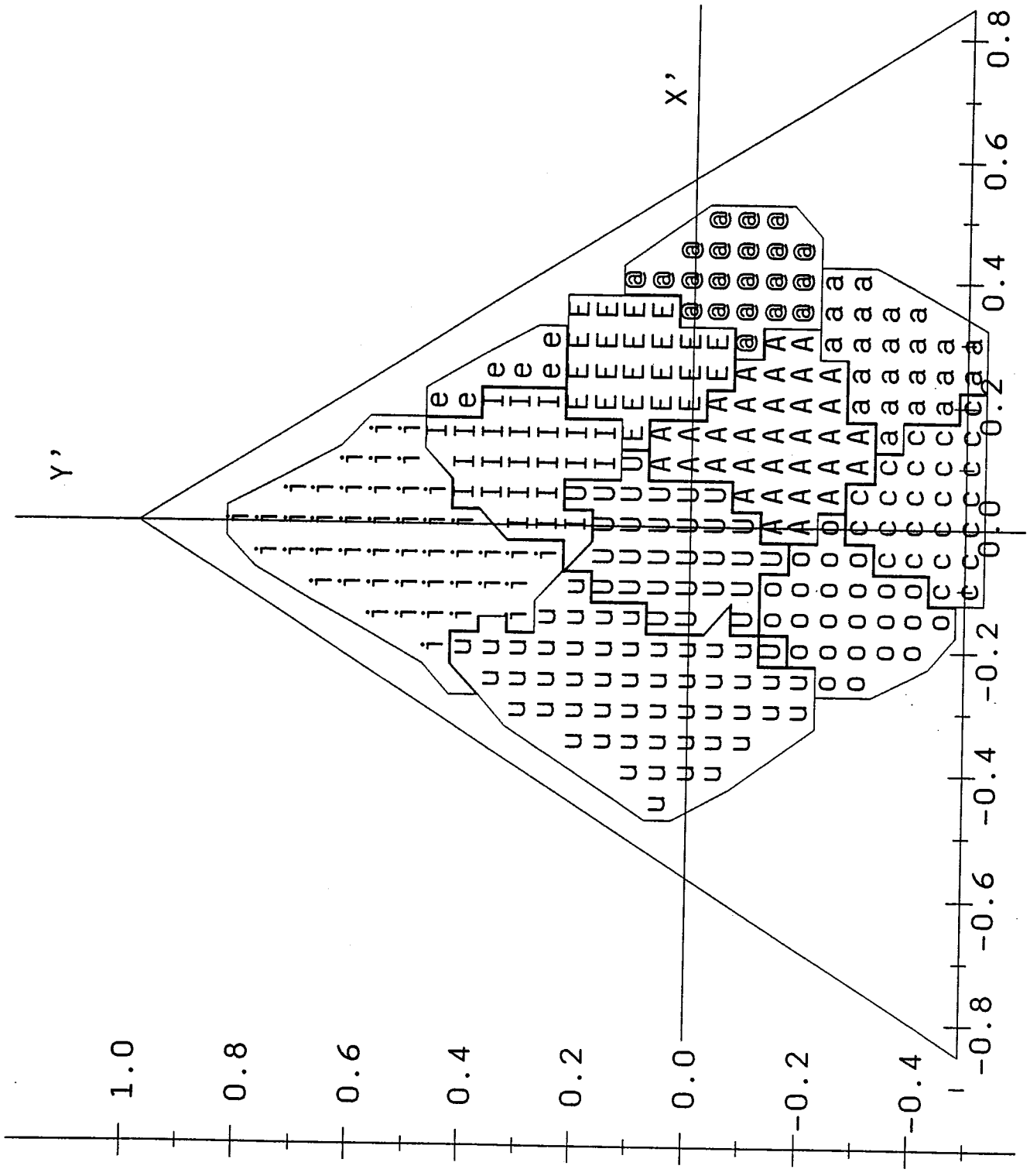


Figure 2-8: (c) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.70$ plane.

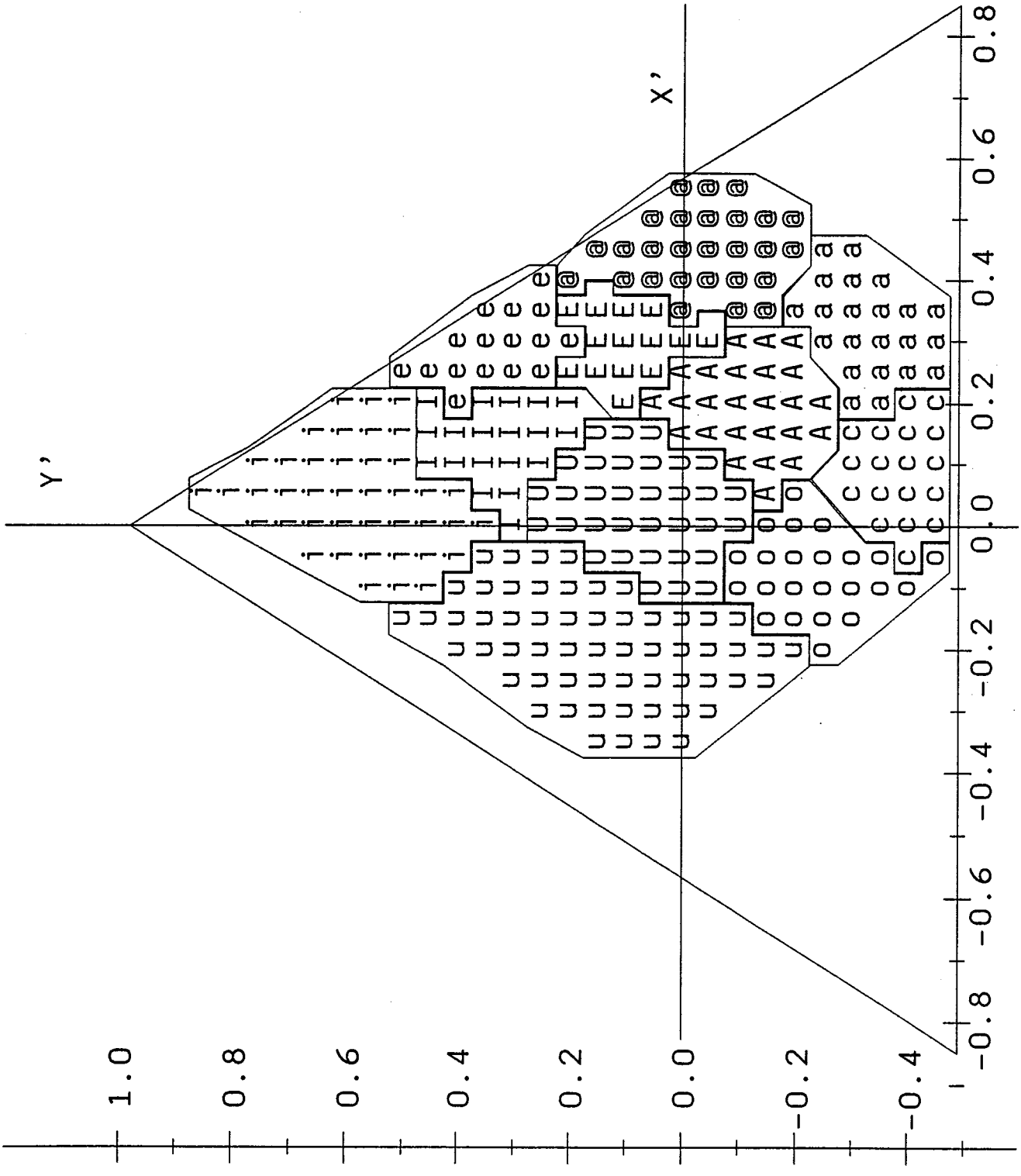


Figure 2-8: (d) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.65$ plane.

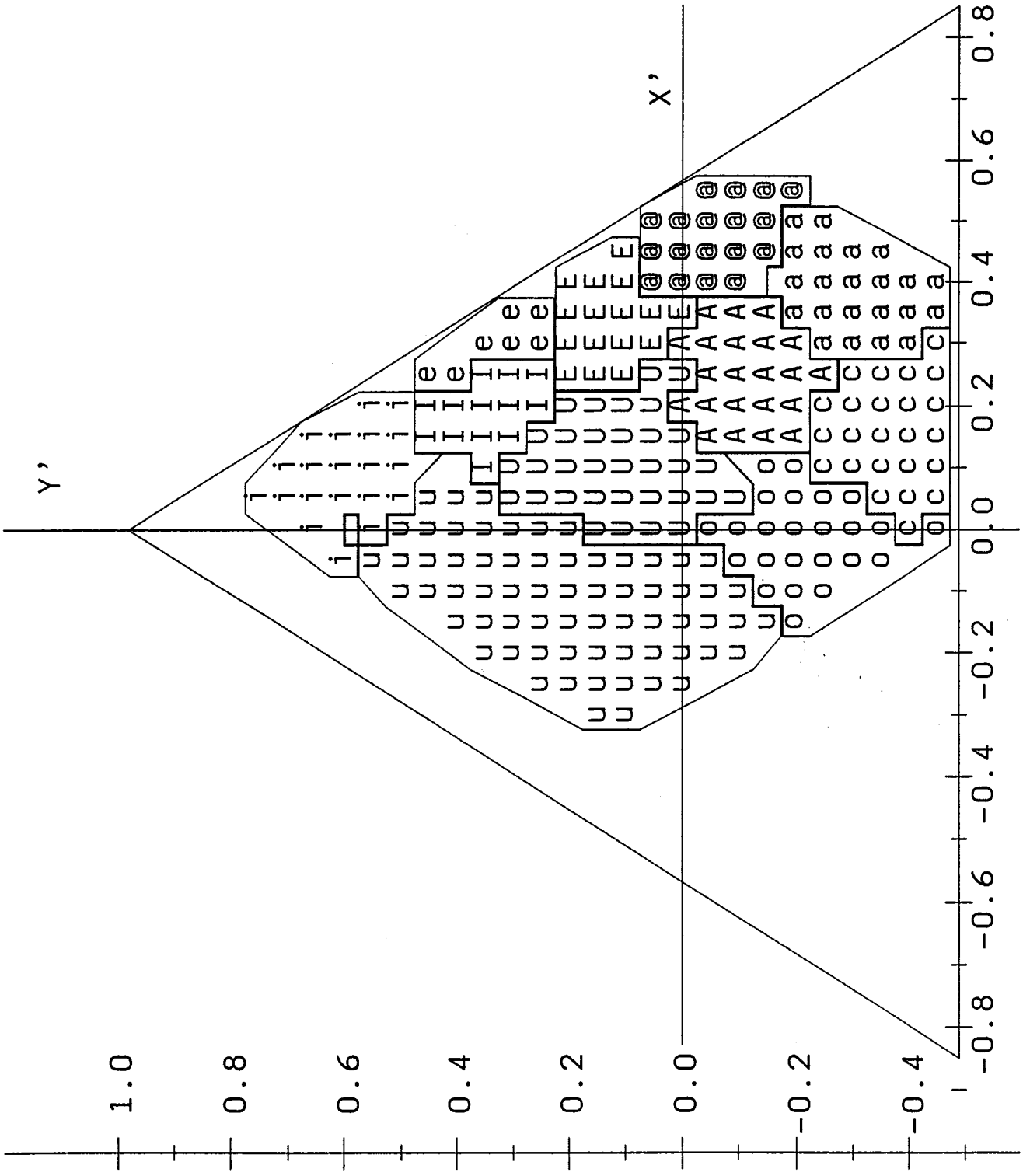


Figure 2-8: (e) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.60$ plane.

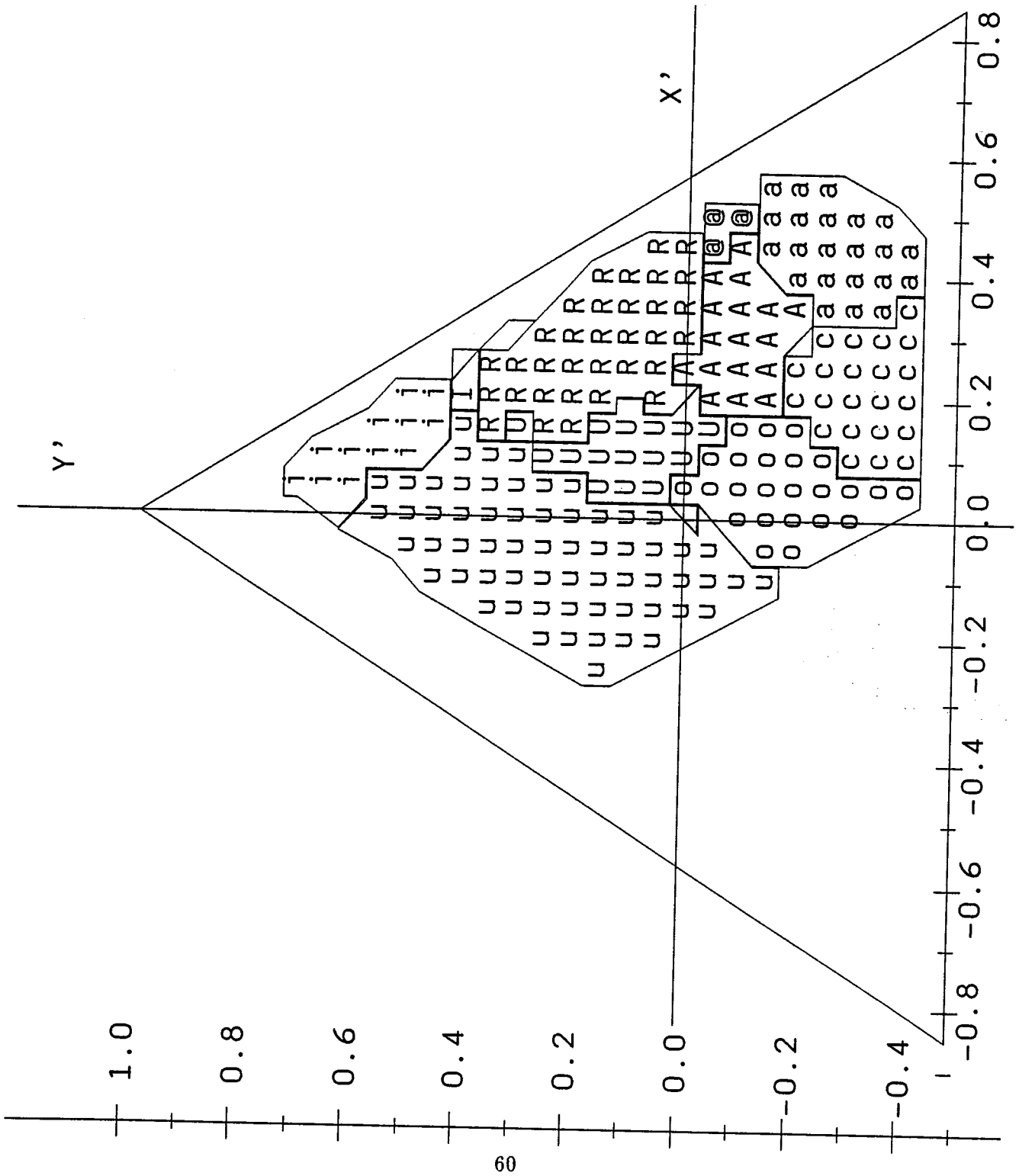


Figure 2-8: (f) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.55$ plane.

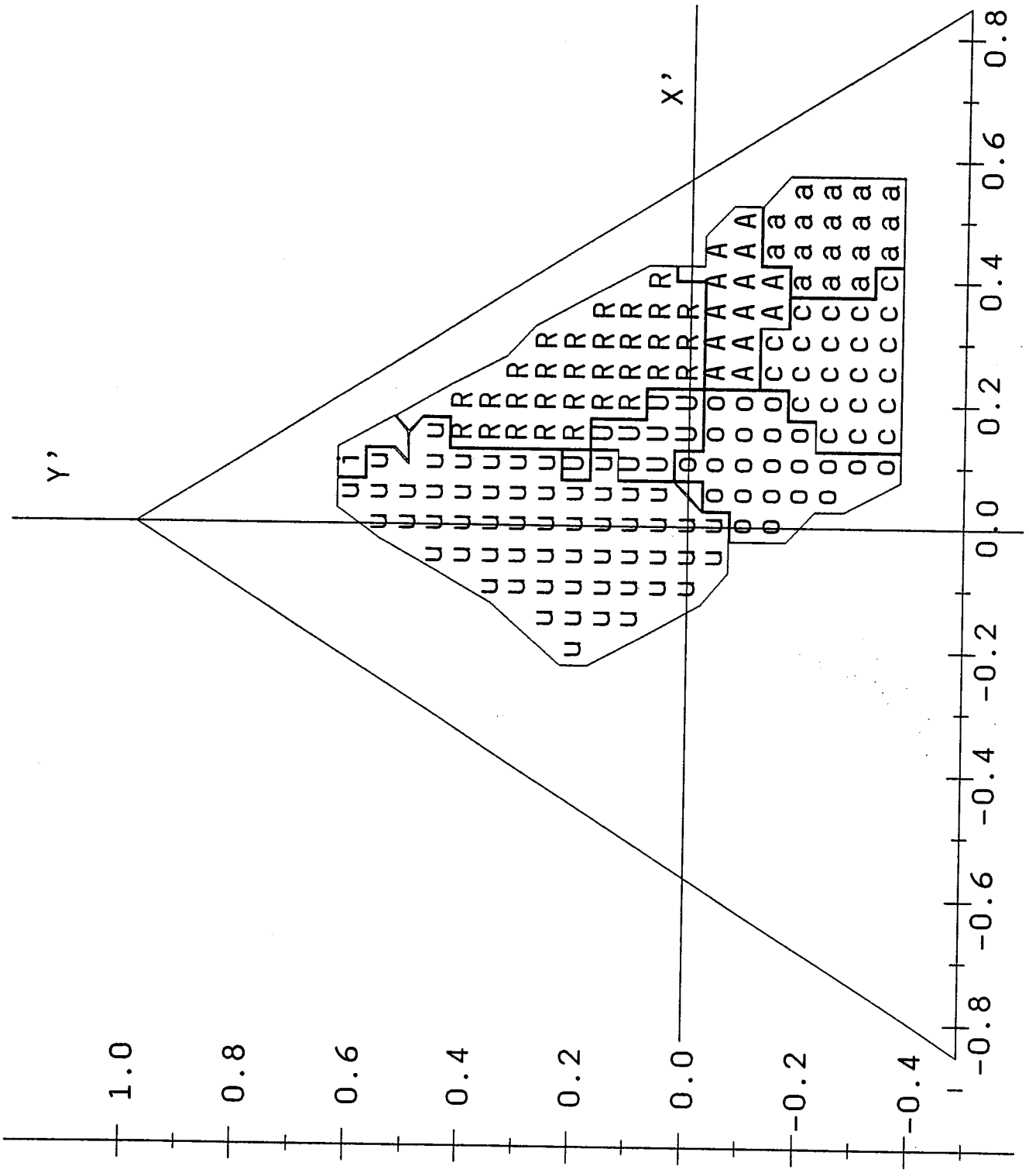
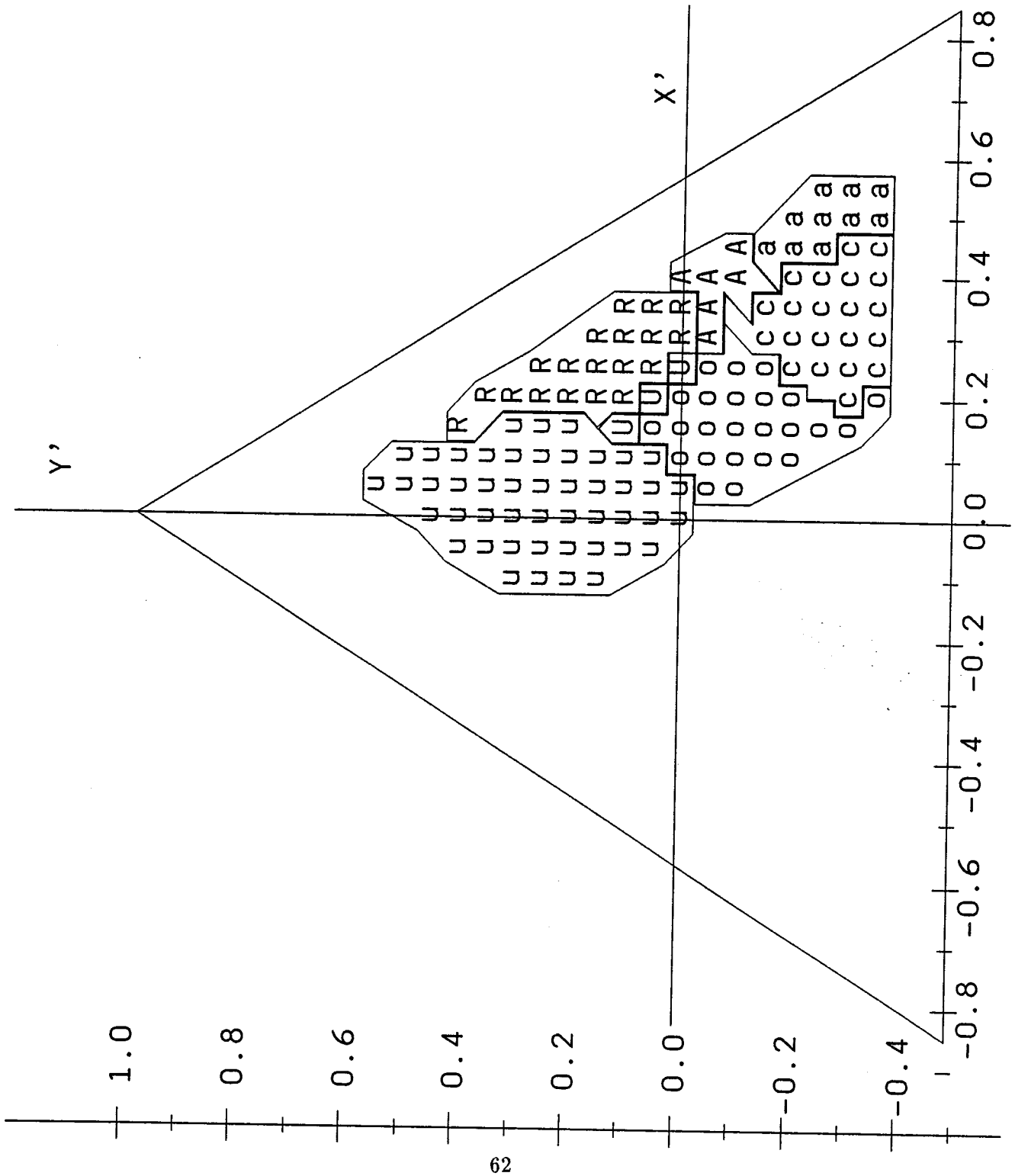


Figure 2-8: (g) Locations of target zones in APS $x'y'$ coordinates constructed on the basis of the plurality identifications from all subject response sets of the synthetic tokens for the $z' = 0.50$ plane.



were approximately equally represented. One tied token (See $x' = 0.1$, $y' = 0.4$, $z' = 0.65$, Fig. 2-8d), could not be represented with a boundary line without causing overlap, so its identification was assigned to one of the two tying categories. It should be noted however that while the cubes associated with the tied points are divided between several identification categories, the exact point locations in the *APS* of these tied tokens do not fall into any zone, but rather, fall between zones.

The two remaining tokens (see Fig. 2-8a) could not be associated with a zone based on the plurality responses without creating overlap in the zones. For these cases, the response category receiving the second highest number of responses for each of these tokens corresponded with an associated zone such that no overlap would be created and was thus chosen as the correct identification for purposes of constructing the zones.

In summary, it is clear that zones for vowel cateogories can be constructed when the identifications for the plurality of subjects are utilized. These zones are adjacent and non-overlapping, and successfully account for over 99% of the synthetic tokens when they are represented as points in the *APS*. In future sections, data analyses pertaining to plurality rates and plurality identifications often will not include the 51 tied or overlapping points discussed above in the data set. These points will be referred to at those times as the "rejected" points.

2.3.5 Qualitative analysis of synthetic speech-based target zones

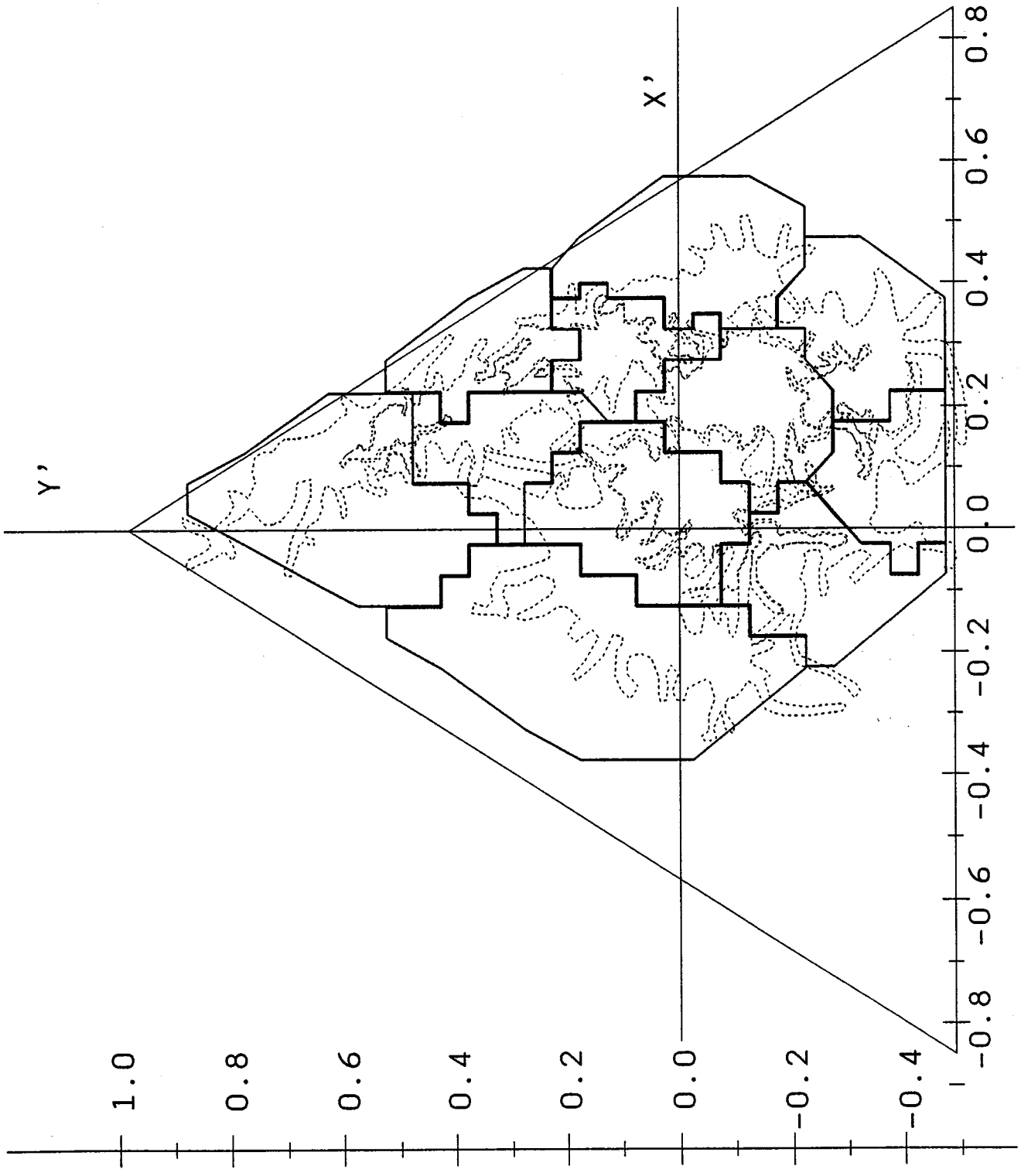
Upon examination of Figures 2-8a-g, it is first apparent that the overall area occupied by the zones is not uniform through z', but changes in shape, shifts in location, and gradually becomes smaller with decreasing values of z'. These changes in overall area reflect the second set of criteria (See Section 2.2.1) utilized in limiting the stimuli to acceptable formant combinations based on production restraints found in natural speech , and therefore, reflect not only these natural constraints, but also define the space available for synthesis. Notice also that the zone boundaries are not stable in terms of the z' dimension, but rather, change with each plane through z'. Additionally, the zones themselves appear to shift with decreasing changes in z' in an upwardly direction and to the right. This apparent shift suggests that like identifications are generally following lines of constant $F1$ and $F2$ back

through z' as is illustrated and discussed in Appendix A.

Figure 2-9 shows the *SSB* zones for the central primary ($z' = 0.70$, from Figure 2-8c) plane overlayed on the most recently estimated zones based on natural-speech productions from Figure 1-9, Chapter 1. The locations of the new *SSB* target zones relative to one another is approximately the same as the locations of the target zones based on natural speech. The locations of the vowel categories /EY/ and /OW/ not used in the natural speech-based (*NSB*) target zones fall predominantly into areas of unclaimed space relative to the *NSB* zones, that is, /EY/ lies to the right of /IH/ and between /IY/ and /EH/, /OW/ lies between /AO/ and /UW/ and is bordered by /AH/ and /UH/. These locations are in agreement with traditional drawings of vowels according to articulatory tongue position. The zone for the retroflex /ER/ lies behind the non-retroflex vowels on the z' axis and ranges from z'= 0.50 to 0.60. At z'= 0.65, the zones for /IH/, /EH/, and /UH/ emerge in the same $x'y'$-space as /ER/, suggesting that the lower z' values are associated with shifting the percept to /ER/. Lower z' values in *APS* translate to lower $F3$ values which have often been associated with perception of retroflexion. The fact that overlap is found between the /ER/ and the /IH/, /EH/, and /UH/ zones when z' is not considered is in general agreement with plots of $F1$ by $F2$ for vowels from Peterson and Barney (1952) which also showed overlap for these categories.

In summarizing this section, we find that the *SSB* zones for vowels in the *APS* change in shape and location as z' changes. These changes reflect the area available for synthesis and the fact that like values of $F1$ and $F2$ are generally associated with the same vowel category, except in the case of perceived retroflexion. The *SSB* zones in the primary planes are in general agreement with the *NSB* zones, with the new *SSB* zones for /EY/ and /OW/ predominantly occupying what was considered unclaimed space with the *NSB* zones. The zone for /ER/ falls in a distinct area behind the zones for /IH/, /EH/, and /UH/, supporting the generally accepted notion that the percept of /ER/ is primarily mediated by a lowered $F3$.

Figure 2-9: Synthetic-speech-based target zones (solid lines) from Figure 2-8 and natural-speech-based target zones (dashed lines) from Figure 1-3 for the $z' = .70$ plane.



2.3.6 Plurality agreements on identifications

The number of subject responses constituting a plurality, the plurality frequency, varied from 5 to 16 out of the 16 possible responses per token. The number of tokens and the percentage of the total number of tokens by plurality frequency are shown in Table 2.5 for all z' planes and for the primary planes only. For example, the row corresponding

Table 2.5: Agreement on identification responses by plurality frequency.

Plurality Frequency	All Planes	%	Primary Planes	%
16/16	320	18.6	170	19.3
15/16	190	11.0	103	11.7
14/16	171	9.9	94	10.7
13/16	148	8.6	75	8.5
12/16	167	9.7	88	10.0
11/16	137	7.9	63	7.1
10/16	145	8.4	77	8.7
9/16	154	8.9	74	8.4
8/16	128	7.4	68	7.7
7/16	71	4.1	34	3.9
6/16	34	2.0	14	1.6
5/16	11	0.6	3	0.3
Total	1676	97.2	863	97.8

to a plurality frequency of 16 indicates the number and percentage of tokens where all subjects' identifications agreed. We find here that all identifications agreed on 320 tokens across all z' planes and 170 tokens in the *primary* planes. Note that, in general, plurality agreements tend to be quite high with almost 20% of the tokens unamimously agreed upon and almost 60% agreed upon by 75% (plurality frequencies of 12 and greater) or more of the identifications. The percentage of tokens for each plurality frequency does not appear to change significantly when limited to tokens only in the primary planes. This suggests that patterns of plurality agreement are relatively unaffected by values of $F3$ outside the normal $F3$ ranges found in natural speech for non-retroflex vowels.

The plurality frequencies for each token (excluding rejected points) are shown in Figures

2-10a-g for each of the seven z' planes, along with the *SSB* zones based on the plurality identifications. Note that the higher plurality frequencies tend to lie to the interior of each zone and smaller frequencies nearer the boundaries. Additionally, comparison of any given target zone across z' planes shows that some z' planes contain generally higher plurality frequencies than others. If plurality frequencies are considered to be representative of the saliency of tokens, with higher values representing greater saliency and lower values representing less saliency, we find that, not only do the target zones vary in their salience relative to one another, but also that a salience gradient may be applied in all three dimensions of each target zone. Such a gradient could be used to estimate the relative likelihood of associating a given point in *APS* with a particular vowel category. Thus a point falling in a region of high salience could be assigned to the vowel category representing that region with relatively high certainty and a point falling in a region of low salience could receive a low certainty identification or multiple, ambiguous identifications.

2.3.7 Confidence ratings

Can confidence ratings be used in addition to plurality frequencies to establish a saliency gradient? Clarke (1960) suggests that confidence ratings may increase the amount of information transmitted an additional 16.5% over identification responses alone. One approach toward answering this question is to determine whether or not confidence ratings are correlated with plurality frequencies. Figures 2-11a-g show the sum of all 16 confidence ratings for each token plotted in the *APS* along with the *SSB* target zones for all seven z' planes. These values have a possible range from 16, were all subjects to assign a token a 1 rating, to 80, were all subjects to assign a token a 5 rating. As with the identification pluralities, we find that, for a given z' plane, larger values generally lie more interior to the target zones and smaller values nearer the boundaries and that values also vary as a group for a single target zone across z' planes. A non-parametric statistical procedure, the Spearman rank-order correlation, was used to measure the degree of association between confidence rating sums and identification plurality numbers. A moderate correlation (R^2= .508) after correcting for ties was found. Figure 2-12 shows the means and standard deviations for the confidence rating sums grouped by plurality number. Thus, while confidence ratings

Figure 2-10: (a) Plurality frequencies (See text) for all tokens in the $z' = 0.80$ plane.

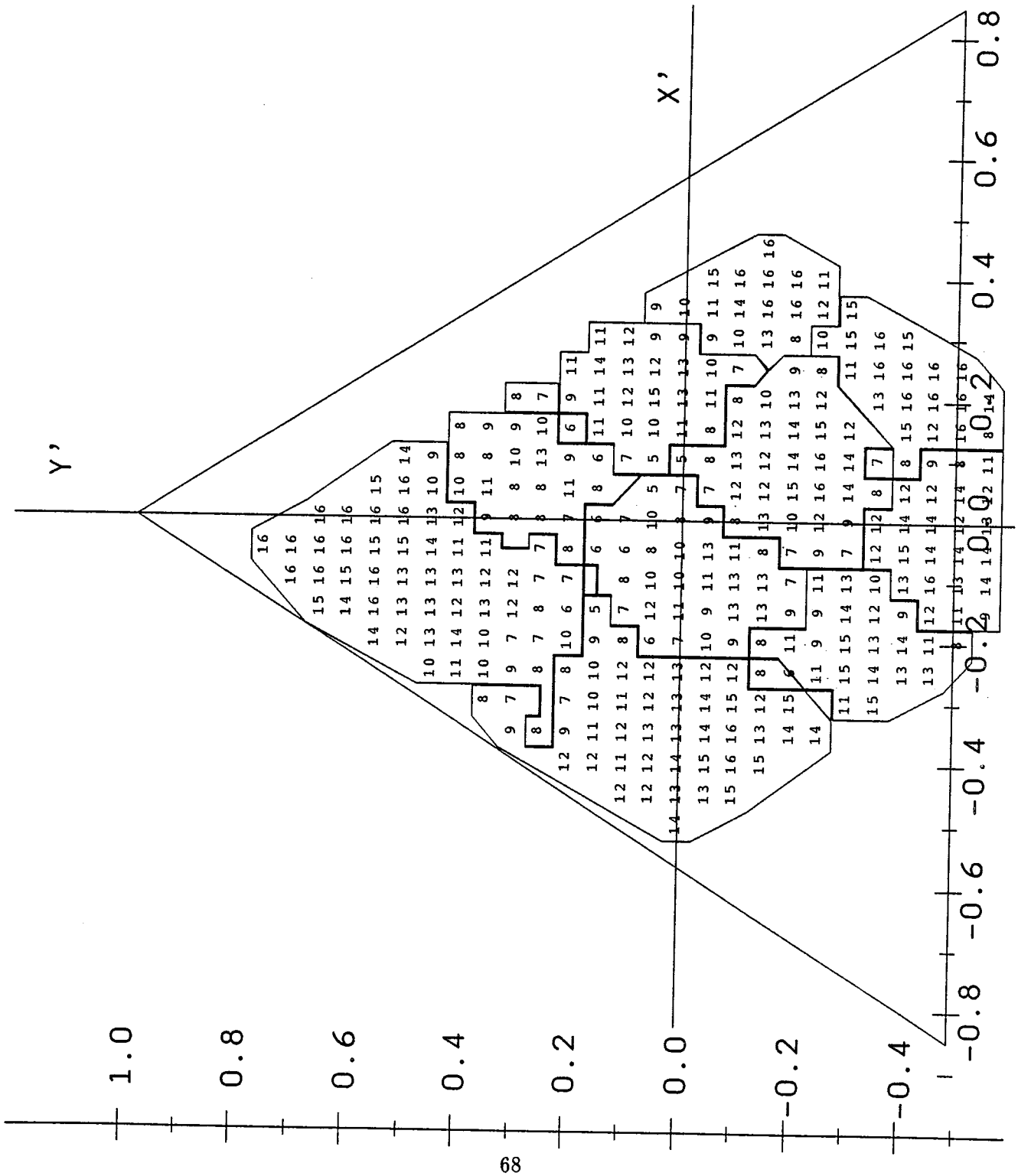


Figure 2-10: (b) Plurality frequencies (See text) for all tokens in the $z' = 0.75$ plane.

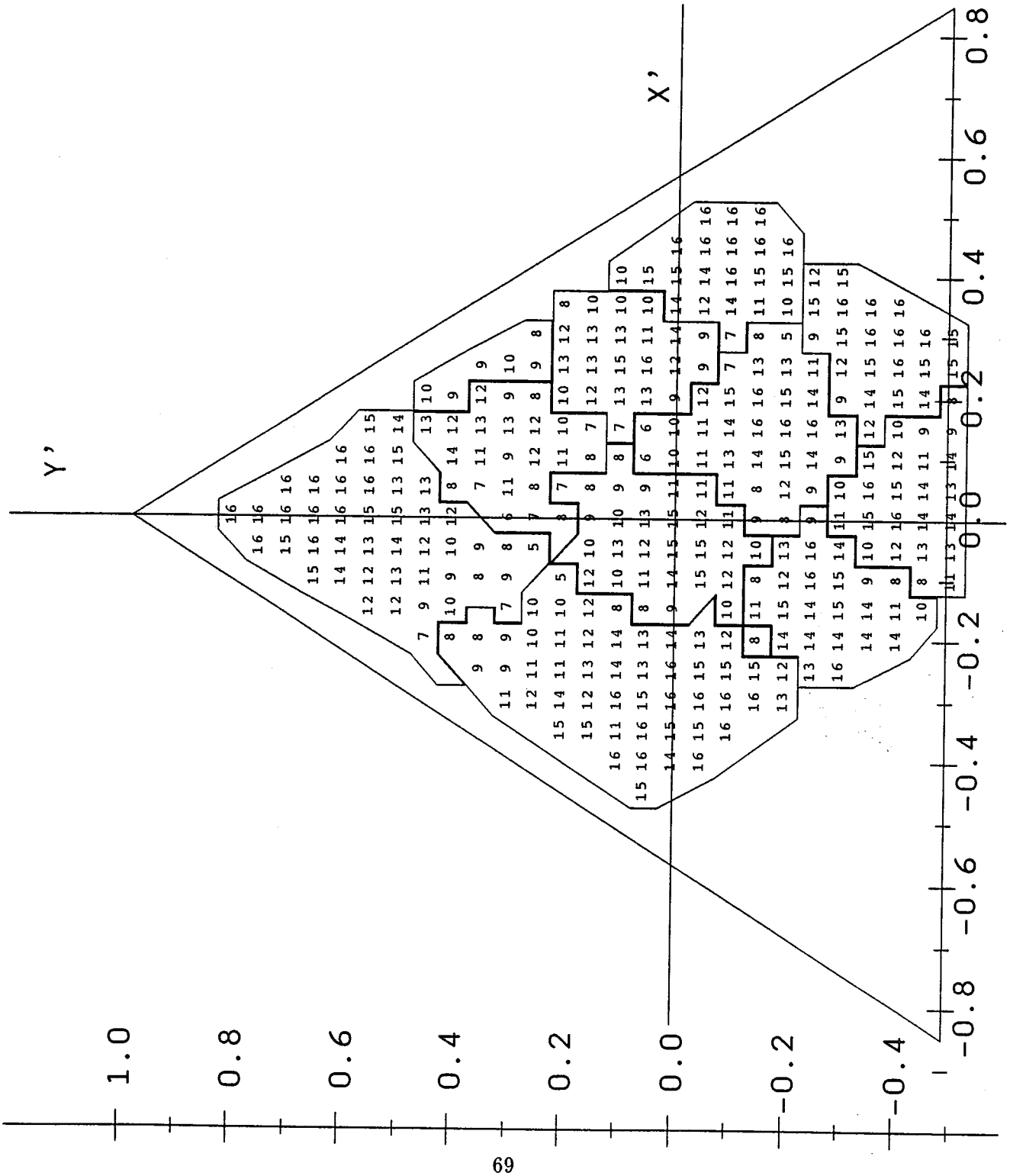


Figure 2-10: (c) Plurality frequencies (See text) for all tokens in the $z' = 0.70$ plane.

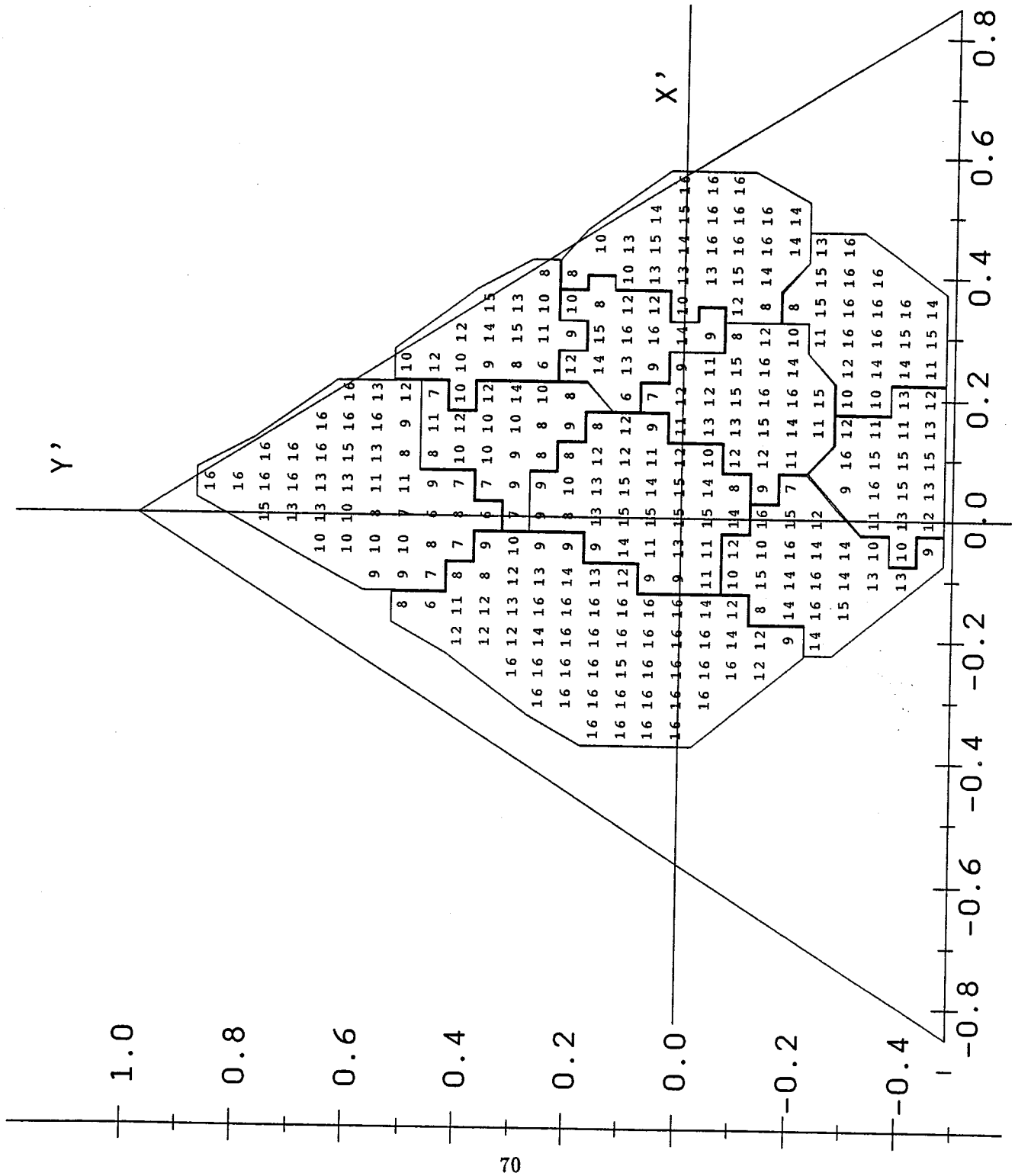


Figure 2-10: (d) Plurality frequencies (See text) for all tokens in the $z' = 0.65$ plane.

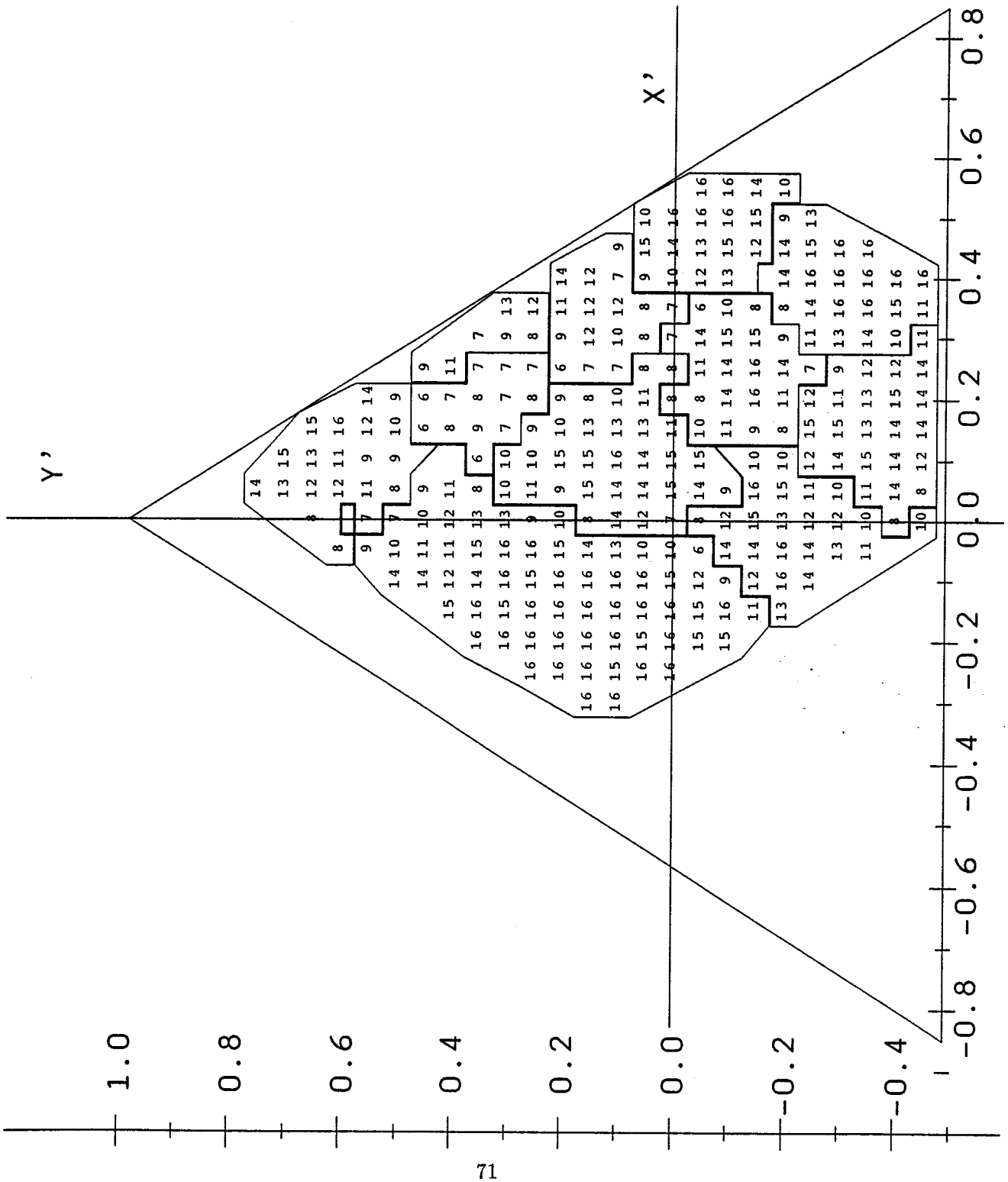


Figure 2-10: (e) Plurality frequencies (See text) for all tokens in the $z' = 0.60$ plane.

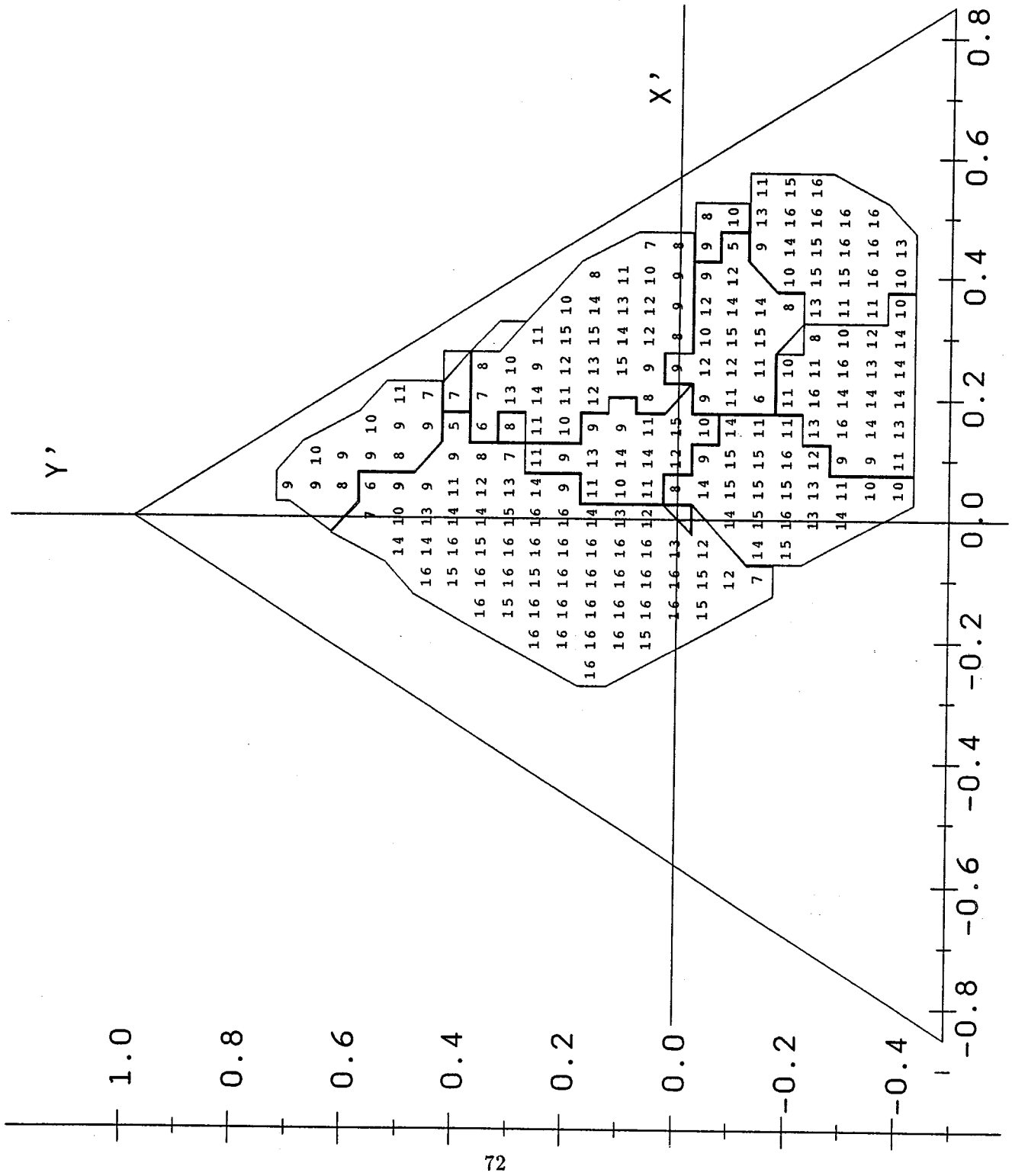


Figure 2-10: (f) Plurality frequencies (See text) for all tokens in the $z' = 0.55$ plane.

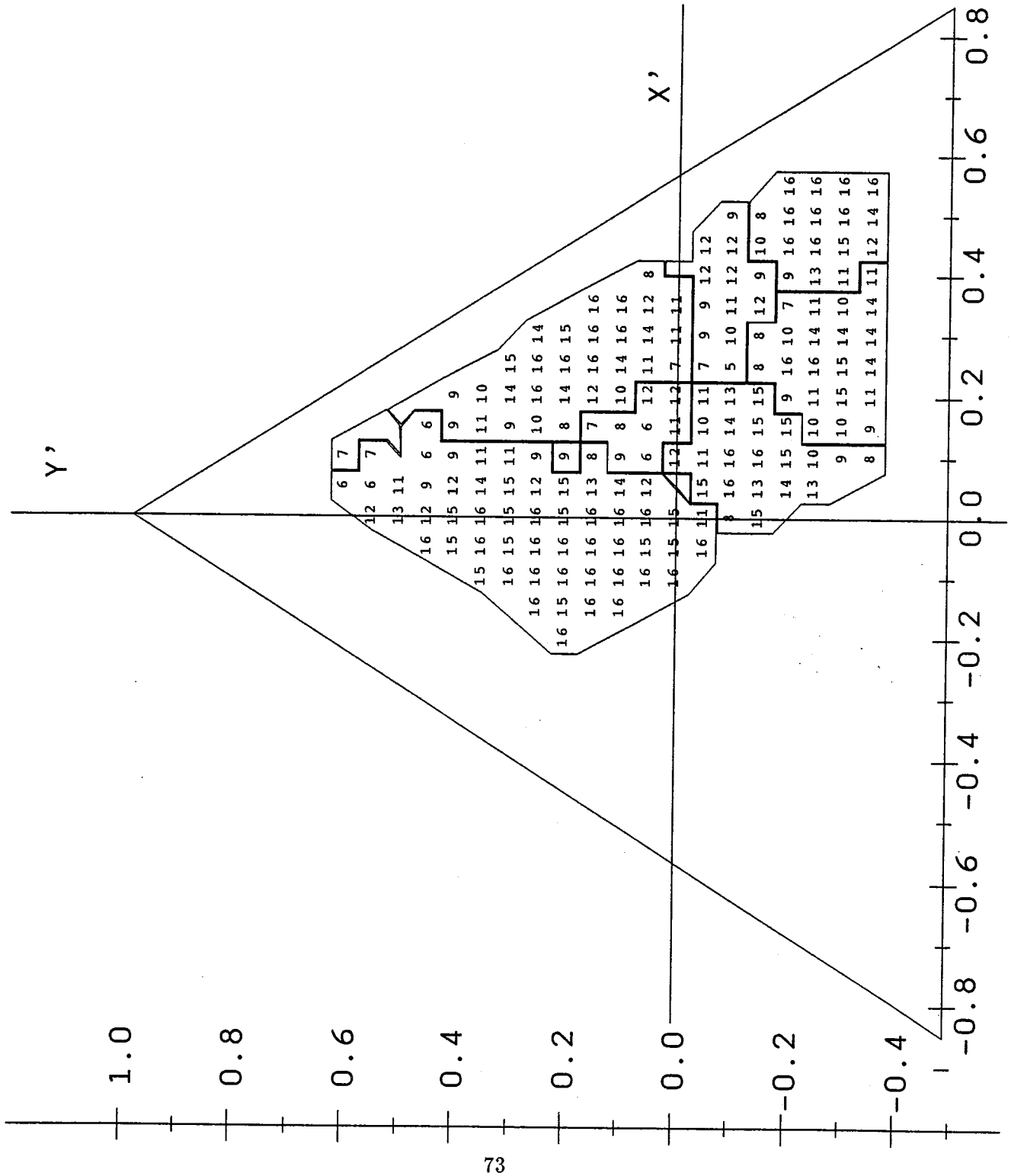
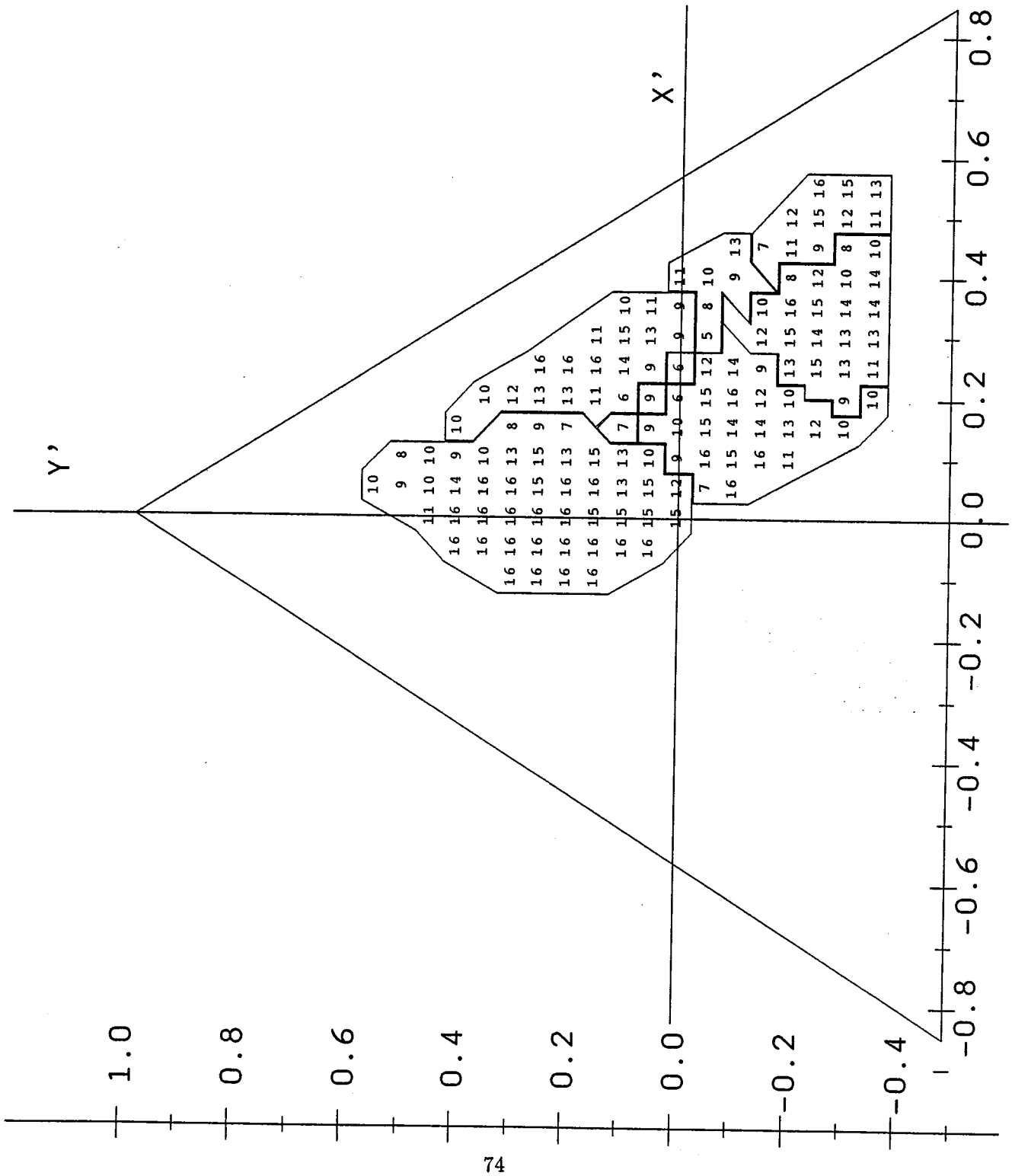


Figure 2-10: (g) Plurality frequencies (See text) for all tokens in the $z' = 0.50$ plane.



are moderately correlated with plurality frequencies, that is, tokens with a greater number of subject responses agreeing on the identification tend to also have higher sums of confidence ratings, the sums of confidence ratings themselves do not appear to substantially add information pertaining to the relative saliency of tokens.

2.3.8 Individual differences in identification responses

As can be noted from Table 2.5, all subject responses agree on only 19.3% of their identifications, or 170 tokens in the primary planes. However, the average within-subject identification agreement, that is, the average of agreements between subjects' first and second response sets, is 75.9% for these planes. This indicates that a subject agrees with himself on the identifications of an average of about 670 tokens. If subject identification consistency is considered an indicator of vowel saliency, then, on the average, 56.6% of the tokens in the primary planes may be considered salient vowels to individual subjects, but yet are classified differently across subjects. In addition, only 304 tokens were shared in common among tokens where within-subjects agreements occurred. Thus, approximately half of the 56.6% represents different tokens to different subjects.

It is acknowledged that some agreements within and between subjects may occur randomly and do not indicate saliency, but rather, add noise to the data. To reduce this noise and pursue this issue further, a third set of identifications for seven of the eight subjects can be considered for one of the primary planes. All subjects (except 1M) classified the tokens in the $z' = 0.70$ plane as part of their initial training. In terms of subject identifications, this may be considered one of the most salient vowel planes in the experiment. Since $F3$ remains constant for any given plane, the response uncertainty for the training set may have been lower than for the actual experiment where tokens from all planes were randomly presented. However, this difference may have been negated by the fact that, at this point in training, the subjects also had less experience with the task.

Percentages of agreement for the identifications from the training response sets and the first and second response sets from the experimental for the $z' = 0.70$ plane were calculated for each of the seven subjects. The average agreement across the seven subjects was 82.7%. The locations of the tokens for which identifications agreed across the three

Figure 2-11: (a) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.80$ plane.

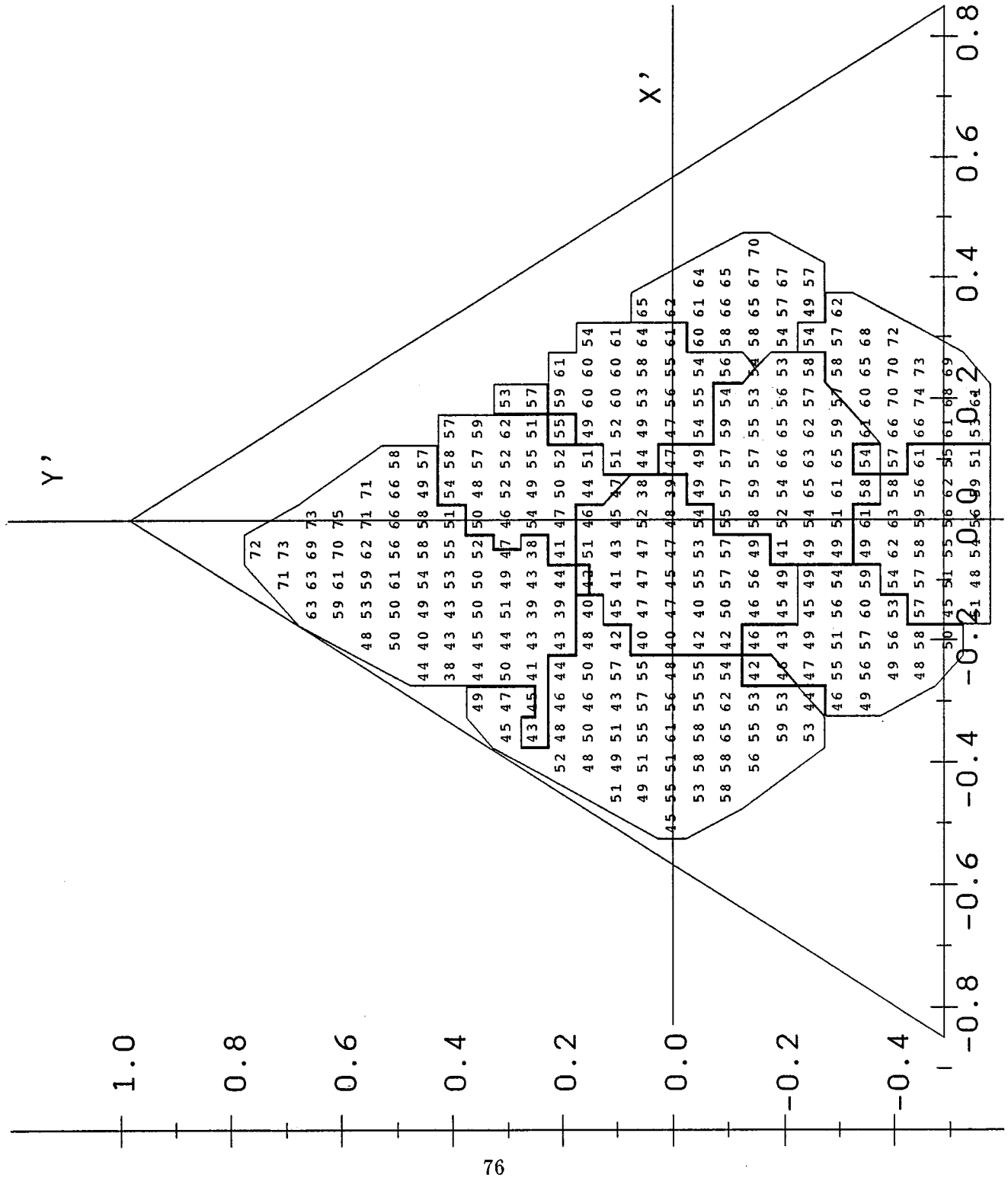


Figure 2-11: (b) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.75$ plane.

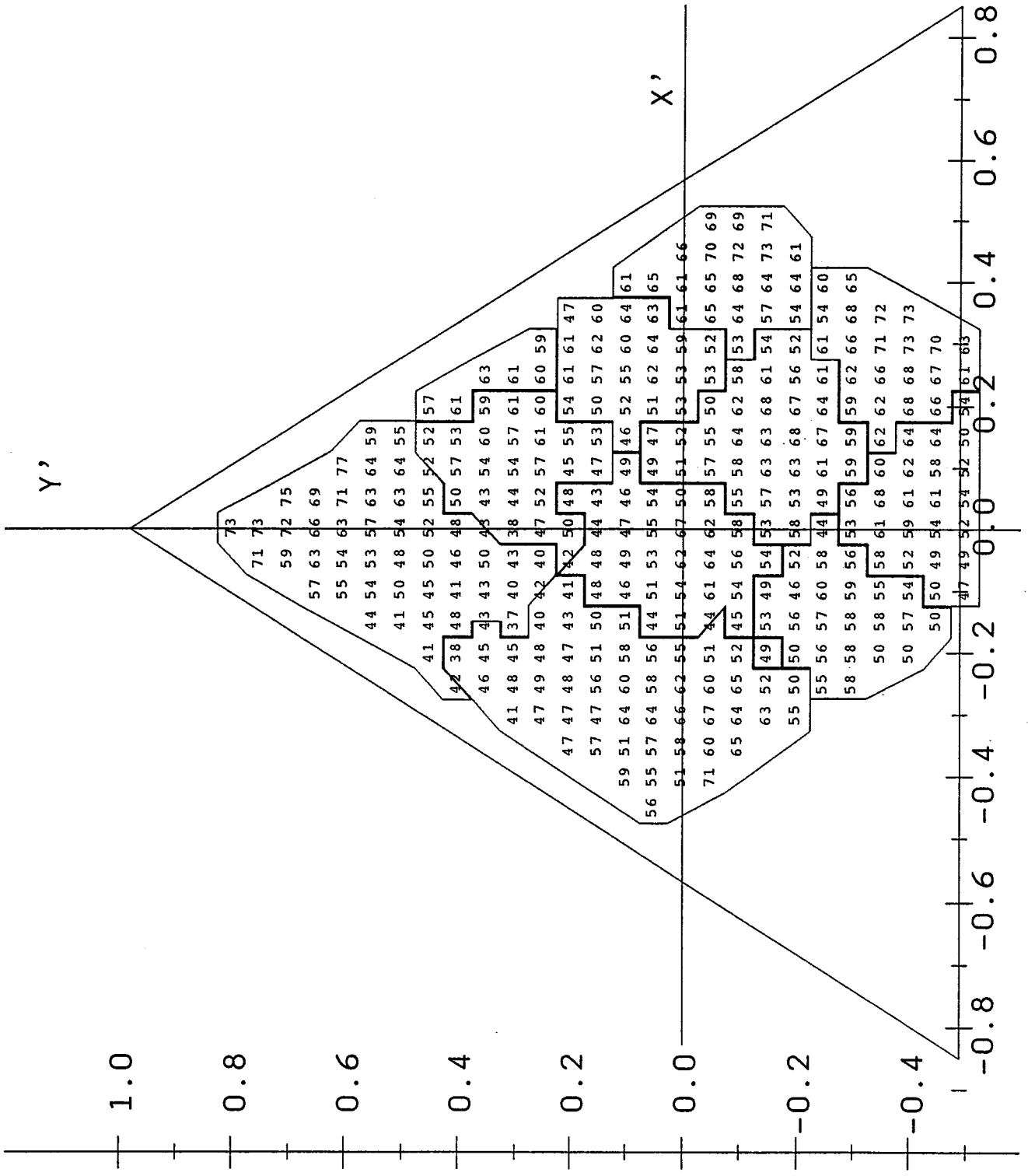


Figure 2-11: (c) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.70$ plane.

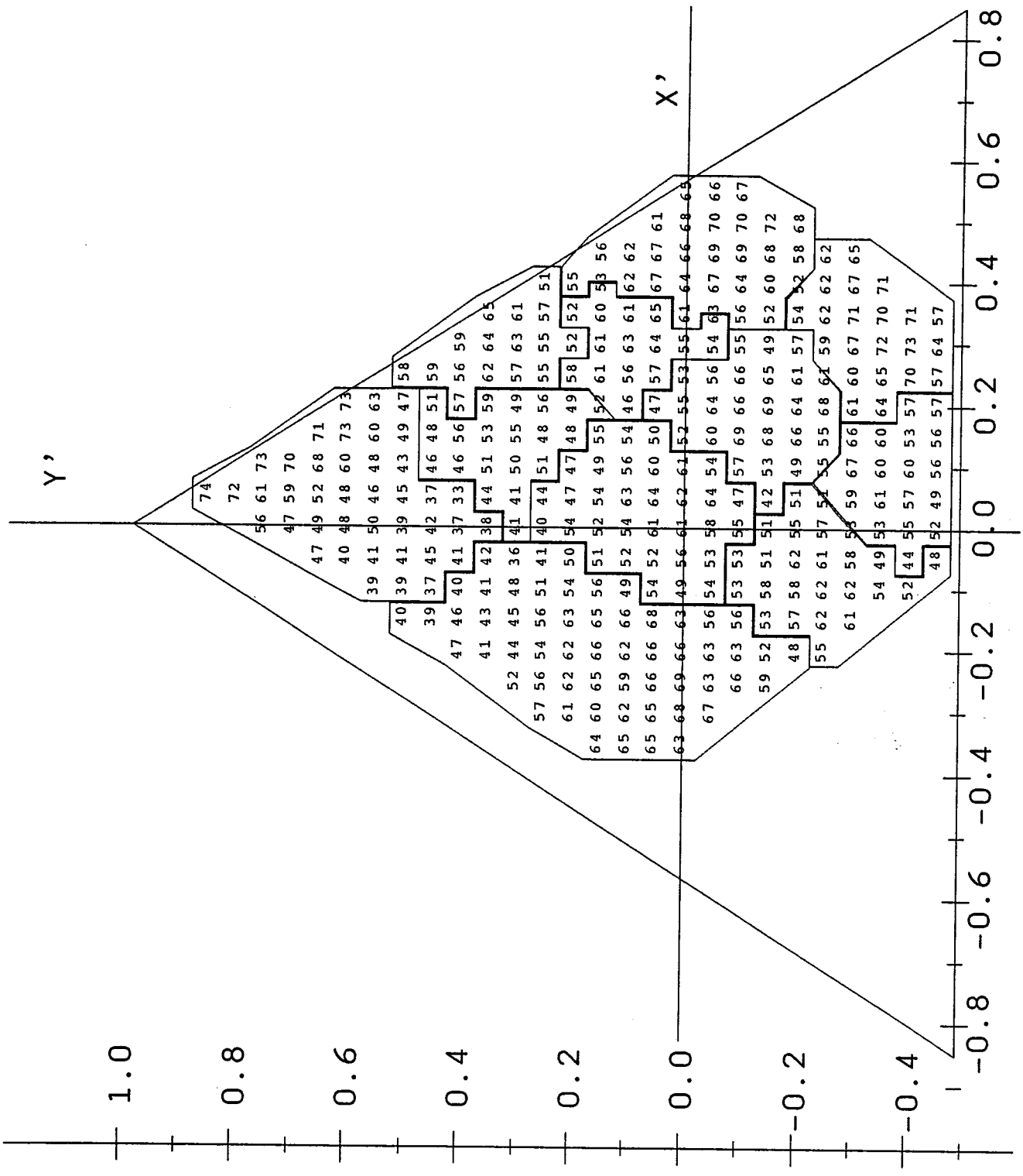


Figure 2-11: (d) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.65$ plane.

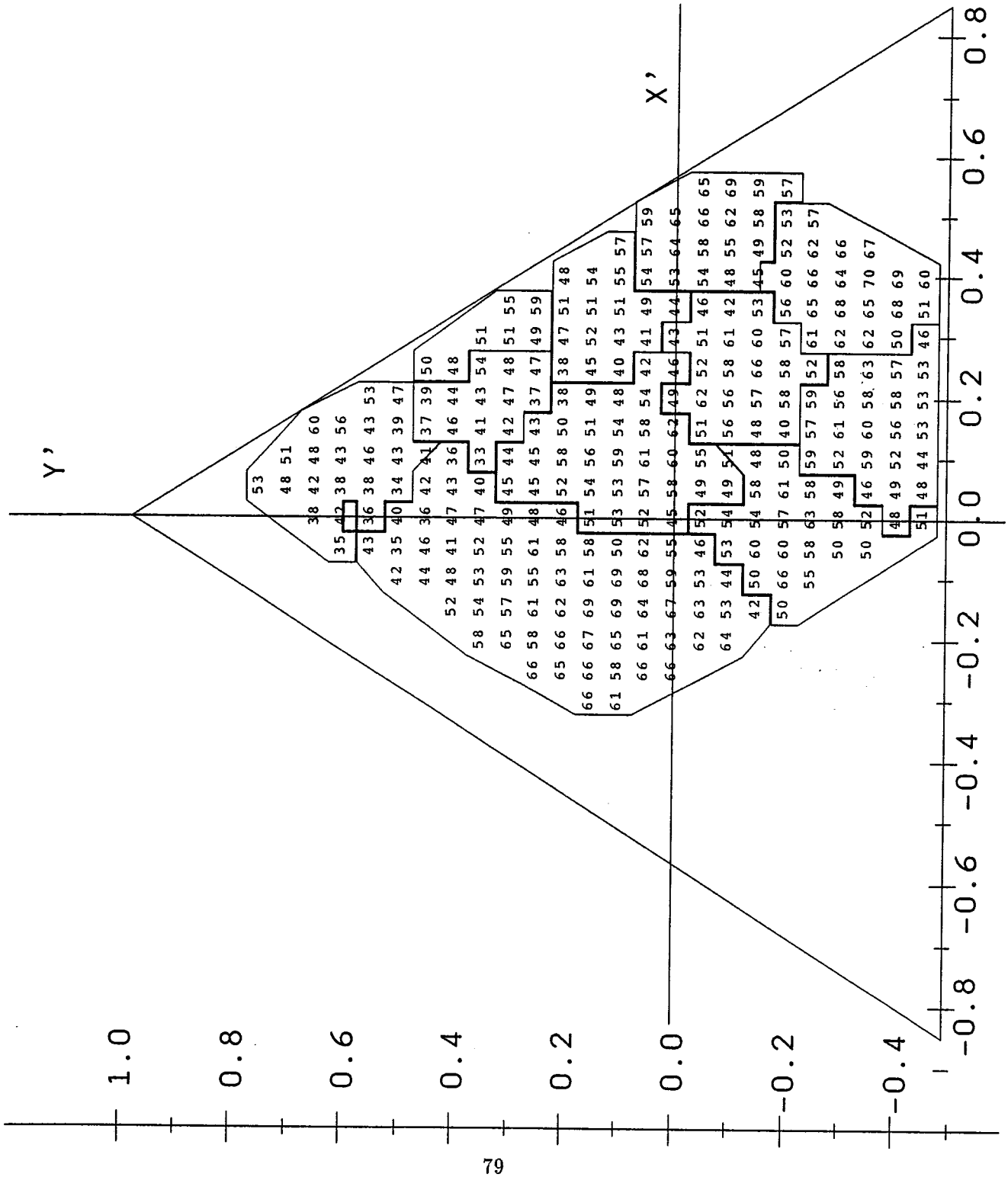


Figure 2-11: (e) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.60$ plane.

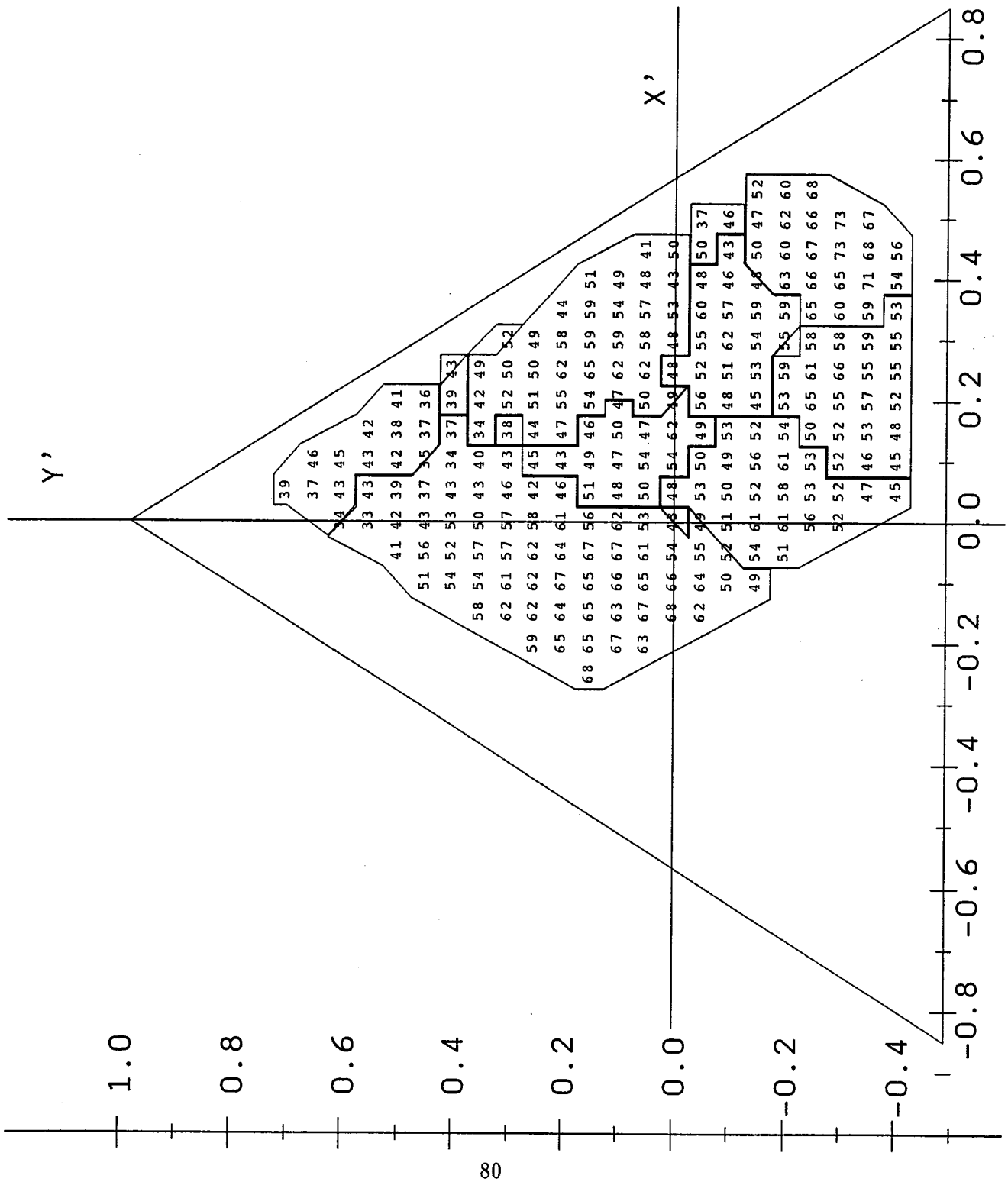


Figure 2-11: (f) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.55$ plane.

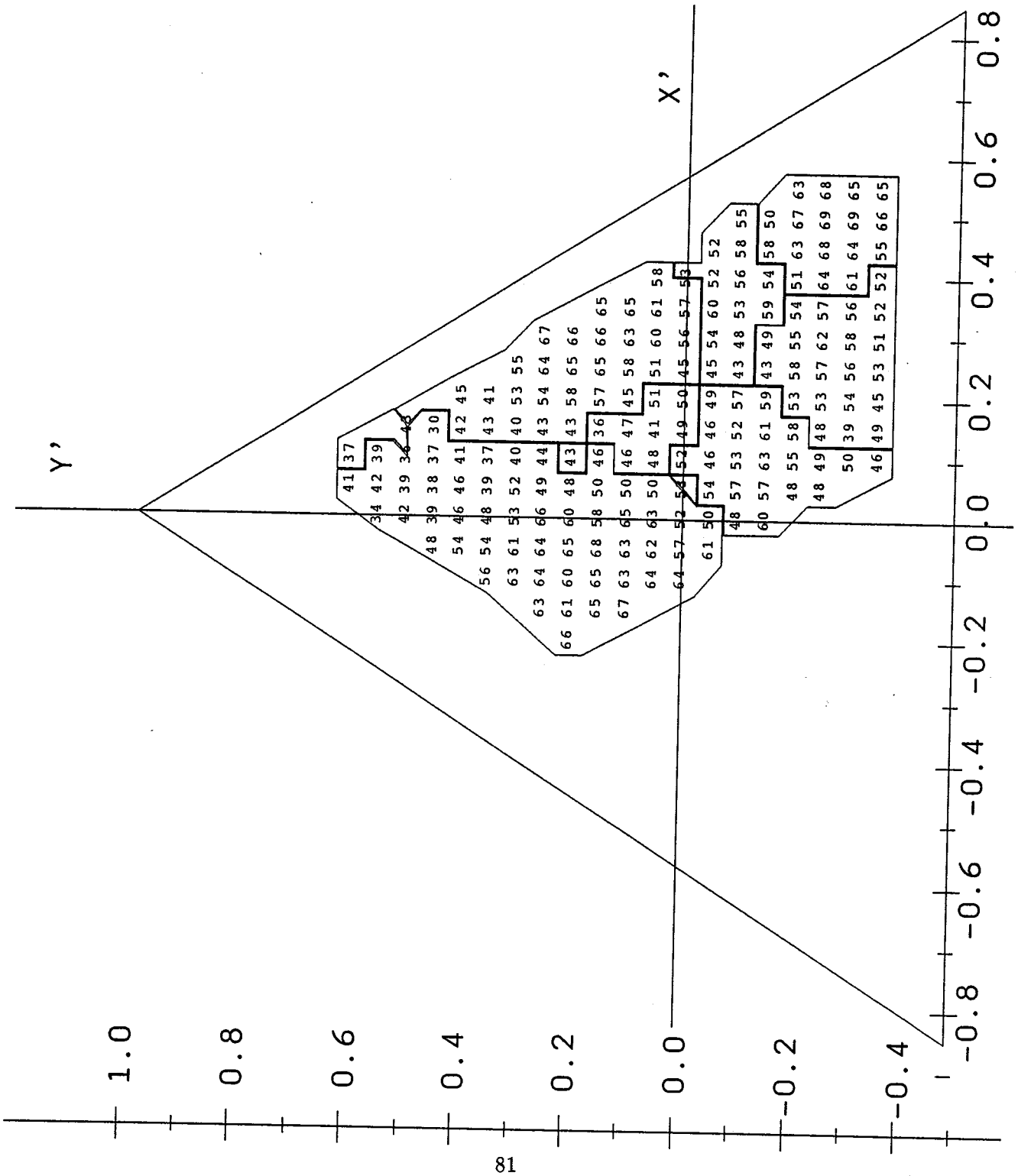


Figure 2-11: (g) Sums of confidence ratings from 16 subject response sets for each token in the $z' = 0.50$ plane.

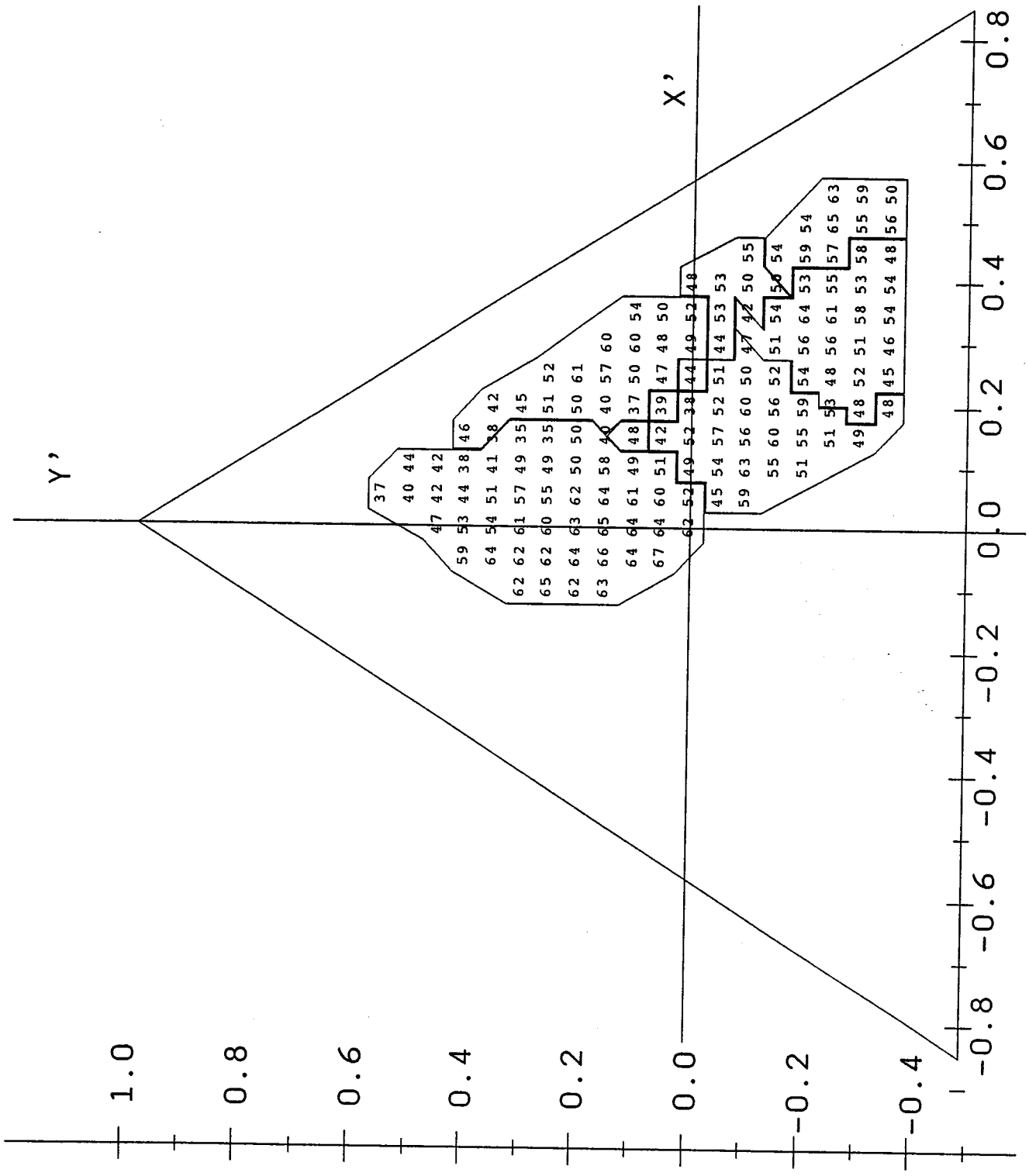
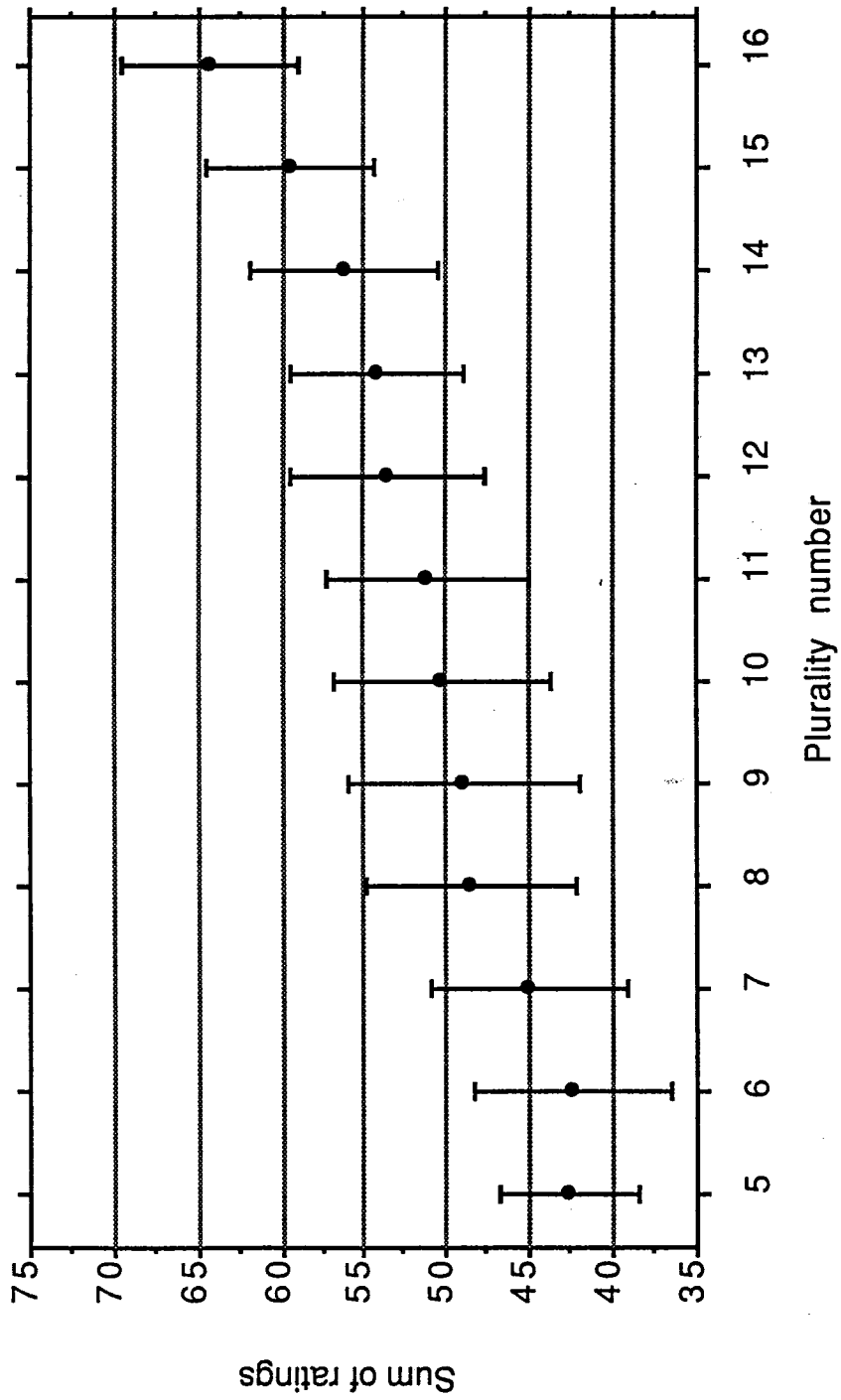


Figure 2-12: Mean sums of confidence ratings for tokens grouped by plurality frequency. Error bars indicate ± 1 standard deviation.



response sets are shown in Figures 2-13a-g for each subject along with the *SSB* target zones for the $z' = 0.70$ plane. In these figures the identifications encircled either disagree with the identifications of the zones in which they fall or fall on boundary lines between zones. These encircled identifications should represent tokens of high saliency to individual subjects, although identified differently or ambiguously by the plurality of all subjects. While a number of the encircled points fall in areas noted previously (see Section 2.2.1) as seemingly difficult for subjects to identify (the areas at the /UW/-/IY/ and /IH/-/EY/ borders) and a few appear to be obvious errors (note the /EY/ responses in the /OW/ and /AO/ zones in Figure 2-13g), the majority fall near or on boundary lines. This analysis suggests that certain formant combinations may produce very salient perceptions of a given vowel quality to individuals which are different from the perceptions of a majority of individuals. Such a suggestion supports the notion that vowel perception may be greatly influenced by individual differences when stimuli are at or near a generalized boundary between two vowel categories. This will potentially add difficulty to defining an accurate general predictive model of vowel perception if based solely on the acoustic attributes of vowels.

Figure 2-13: (a) Locations of tokens for which identifications agreed across three response sets for subject 1F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

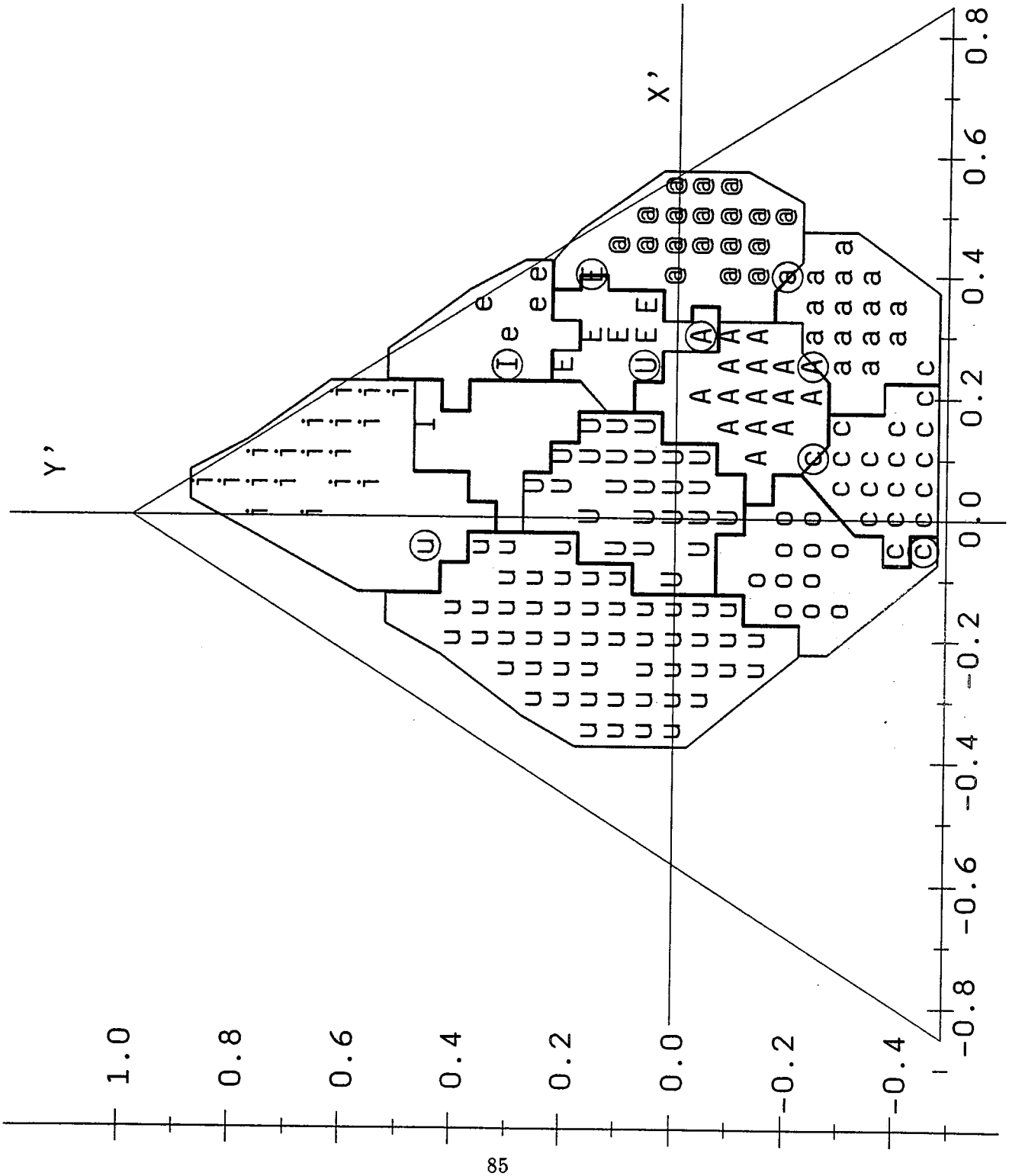


Figure 2-13: (b) Locations of tokens for which identifications agreed across three response sets for subject 2F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

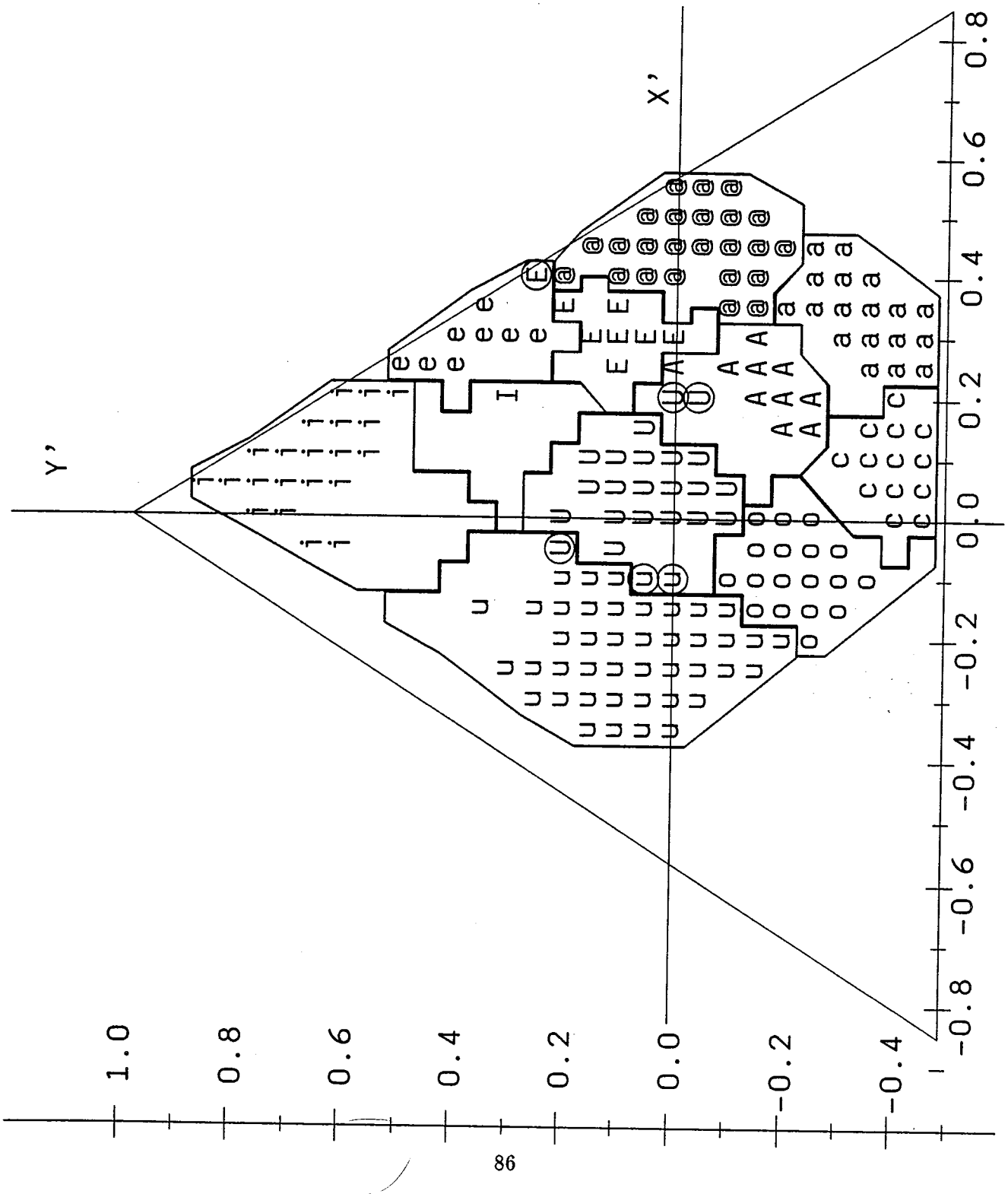


Figure 2-13: (c) Locations of tokens for which identifications agreed across three response sets for subject 3F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

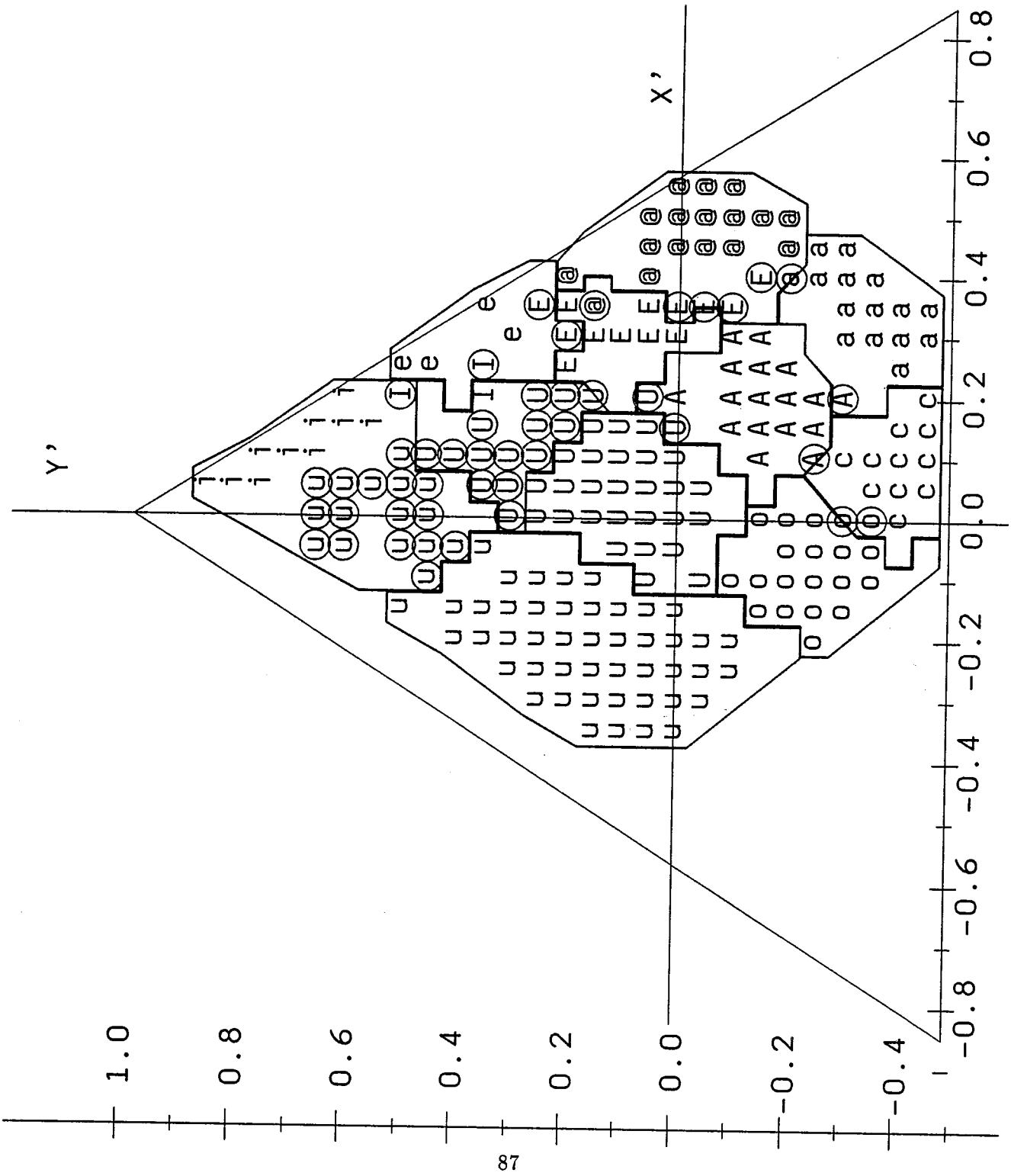


Figure 2-13: (d) Locations of tokens for which identifications agreed across three response sets for subject 5F in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

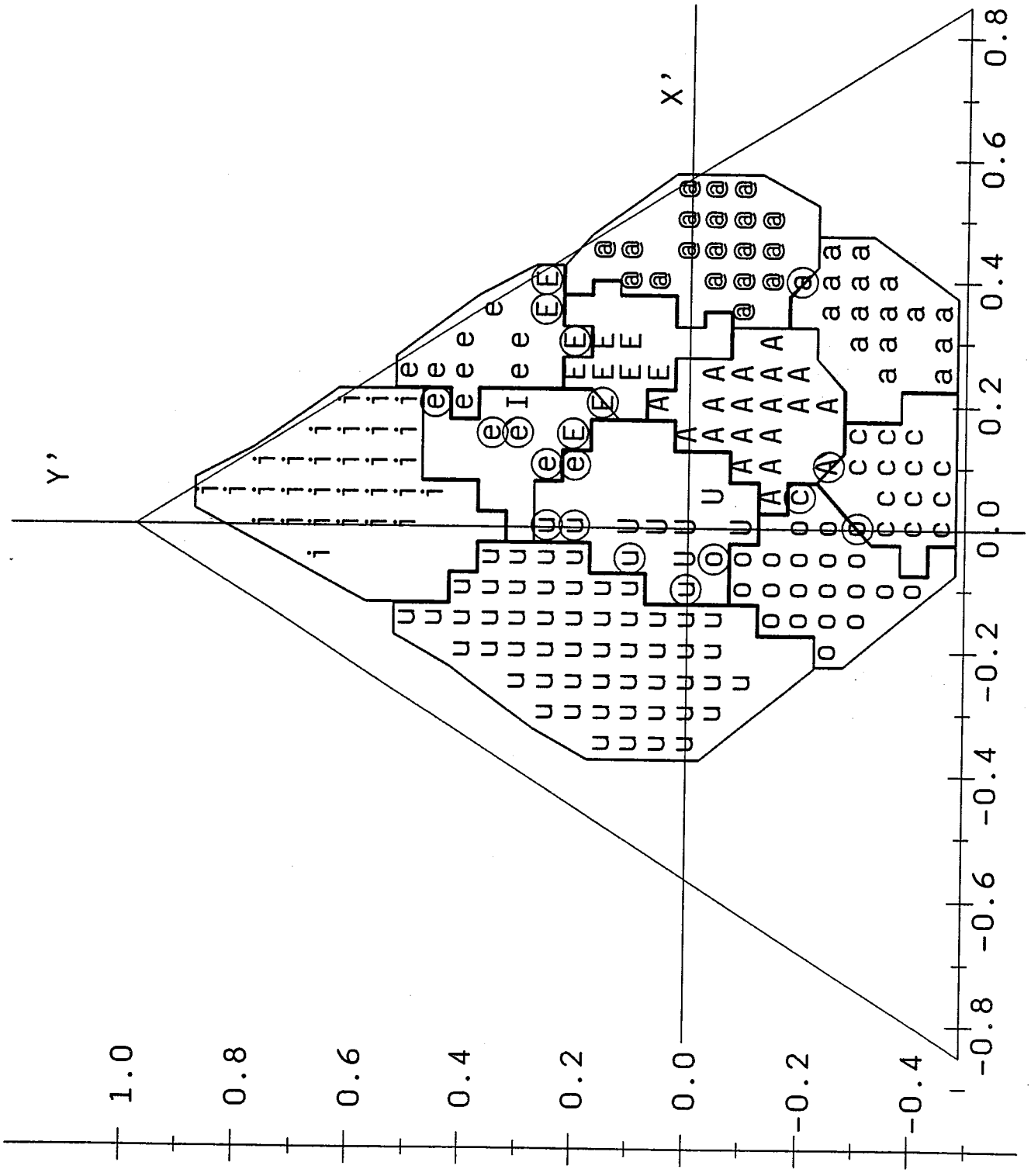


Figure 2-13: (e) Locations of tokens for which identifications agreed across three response sets for subject 3M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

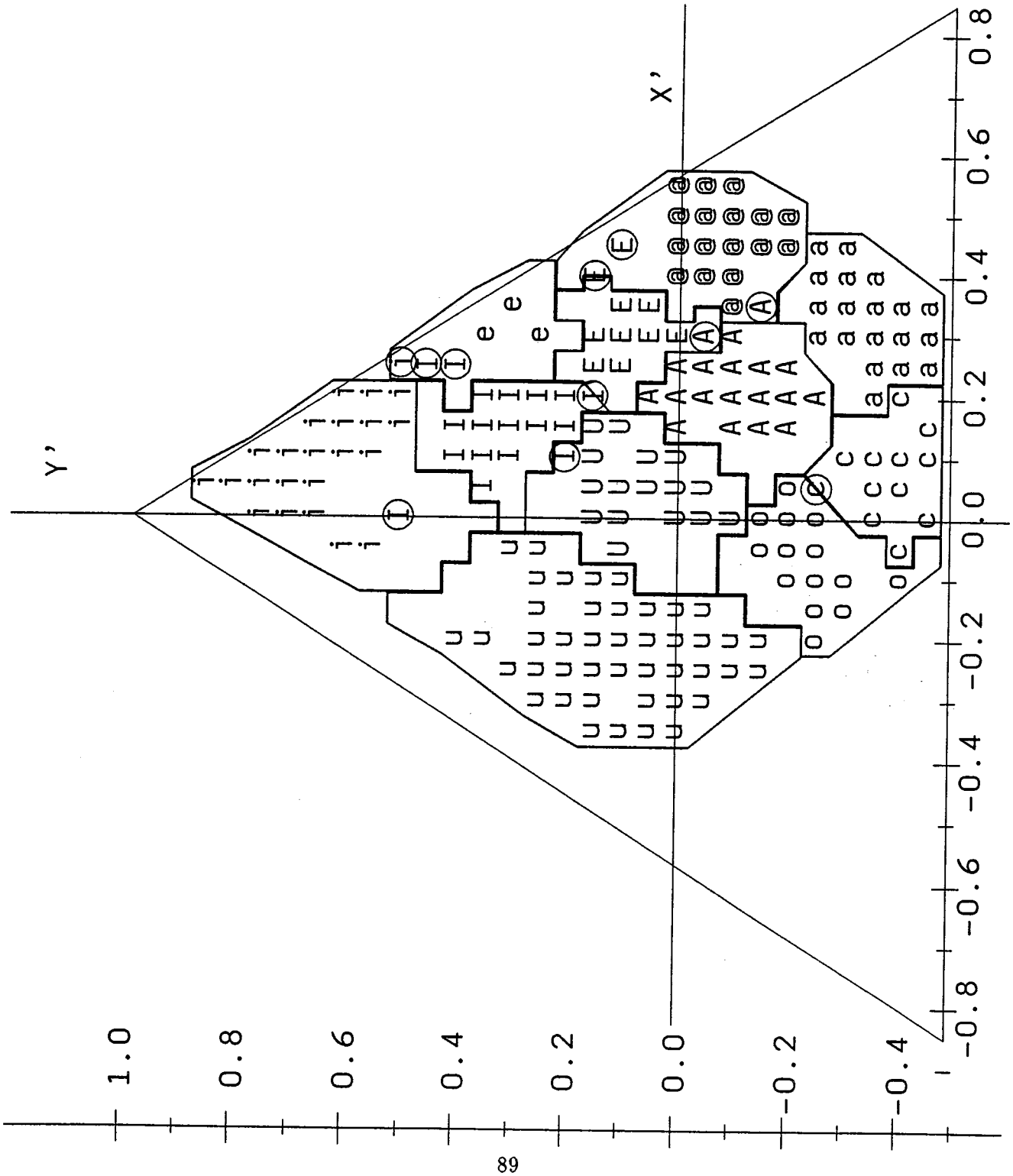


Figure 2-13: (f) Locations of tokens for which identifications agreed across three response sets for subject 4M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.

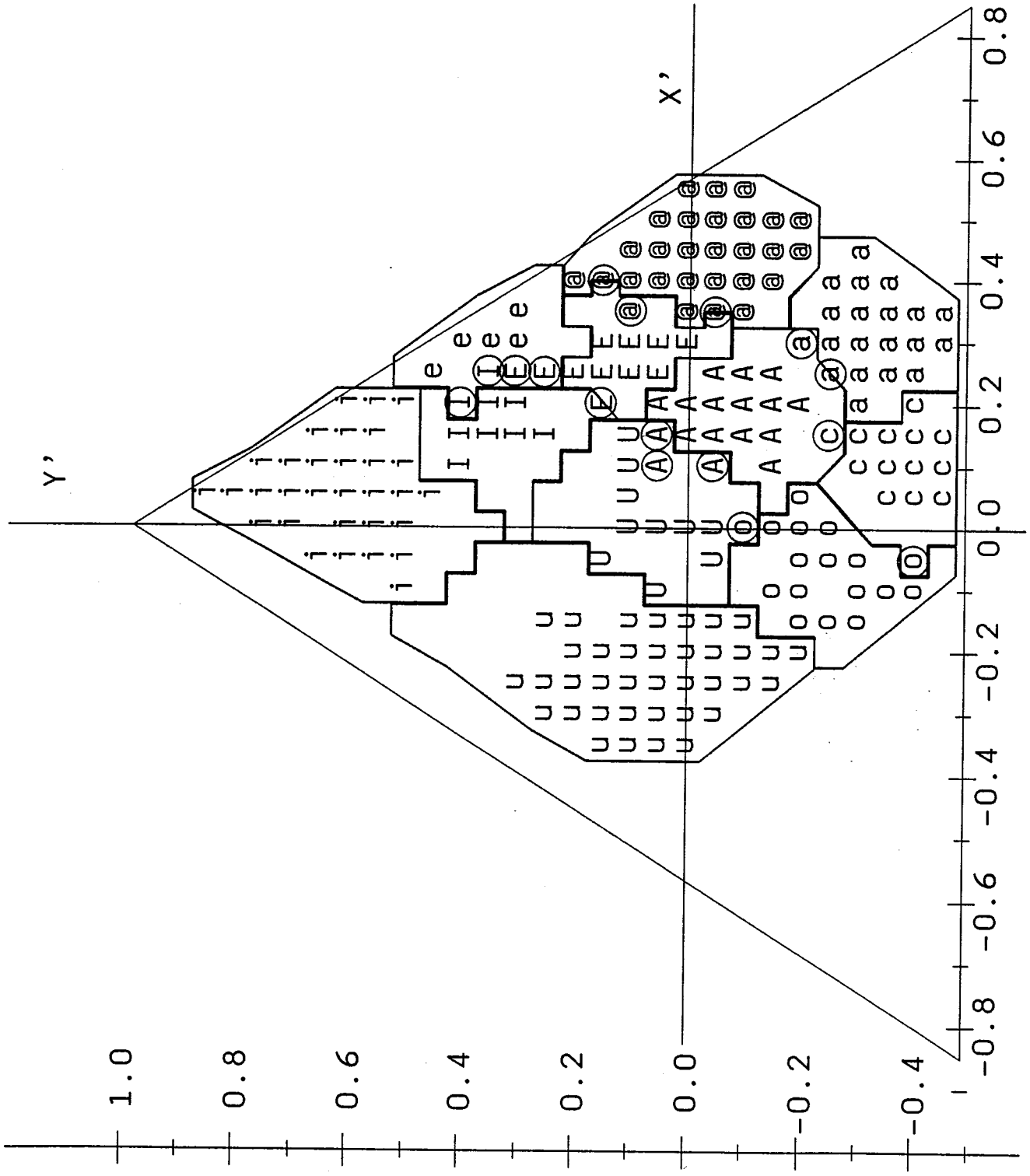
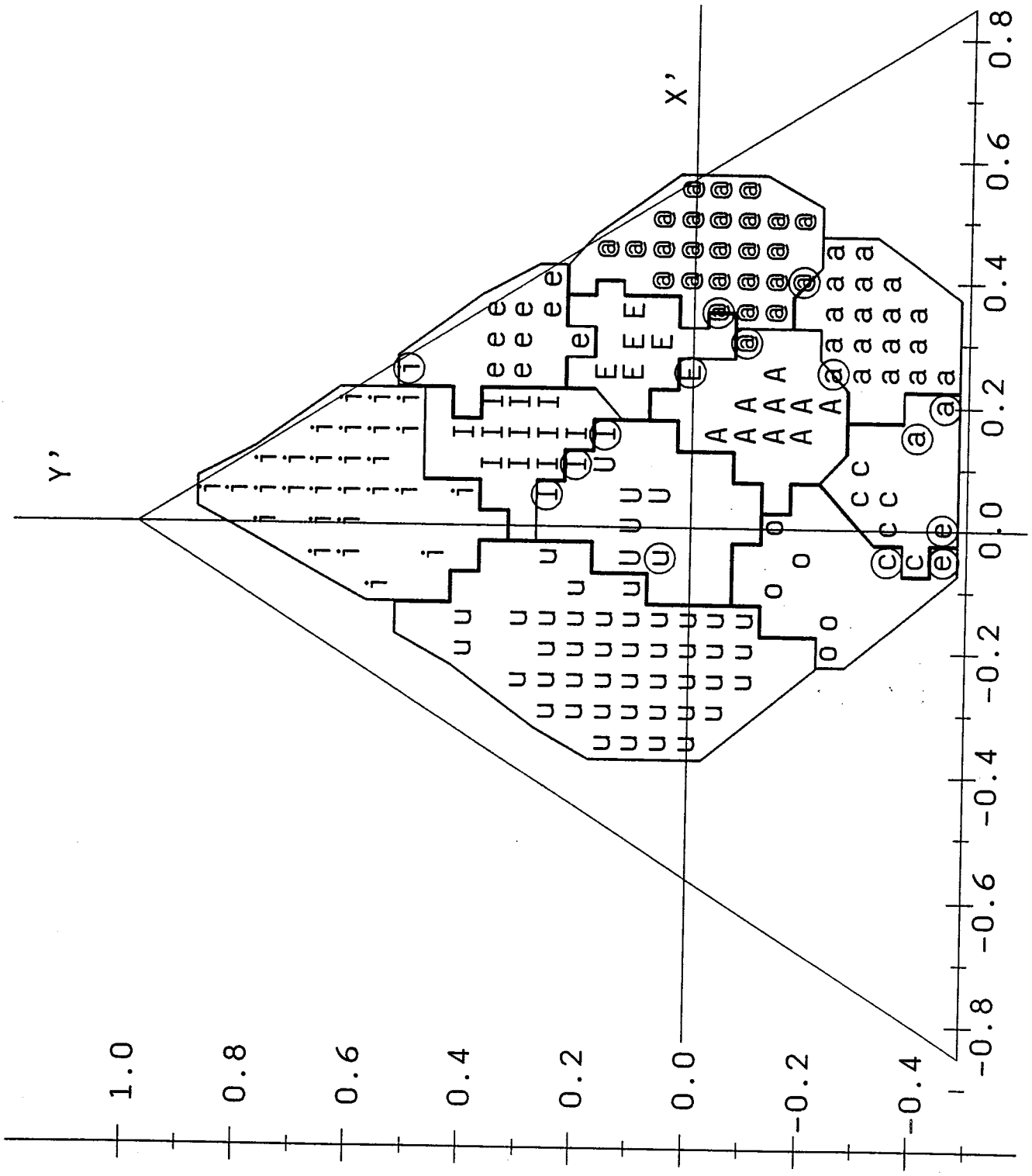


Figure 2-13: (g) Locations of tokens for which identifications agreed across three response sets for subject 5M in the $z' = .70$ plane. Encircled points indicate disagreements or ambiguous locations as compared to plurality identifications.



2.3.9 Linear Discriminant Analyses

To investigate the relation between the plurality identifications of the synthetic vowel tokens and their acoustic variables, a series of linear discriminant function analyses (Tatsuoka, 1970) were performed on the data. This analysis technique has often been utilized in past speech research (for examples, see Neary, 1977; Assmann, Neary, and Hogan, 1982; Syrdal and Gopal, 1986) to provide indices of resolution. These analyses perform a statistical classification function utilizing continuous variables along one or several dimensions, and base the classification on an *a posteriori* probability (*APP*) of group membership assuming multivariate normal distributions of the variables about group means. Results will be presented here using the R (resubstitution) method of classification whereby each token is classified according to calculations based on all the data. The percent correct classification calculated by this method provides a relative index of resolution for how well a given data set may be divided into groups, and the average *APP*, an index of the relative strength of group membership, where group means become more widely separated and group data more closely clustered as the average *APP* approaches one. Thus, linear discriminant analysis provides a method for quantitatively describing how a set of variables may classify a given set of data in a linear statistical fashion and will be used here to evaluate how traditional vowel measures (i.e., $F1$, $F2$, and $F3$), as well as *APS* coordinate values, account for the plurality identifications.

Table 2.6 shows the percent correct and average *APP* scores from a series of linear discriminant function analyses on the 1674 plurality identifications (excluding rejected points) with respect to various combinations of $F1$, $F2$, and $F3$; x, y, and z; and x', y', and z'. It is apparent from the table that classification by $F1$ and $F2$ is quite accurate, but that classification accuracy does increase when $F3$ is included. Classification by log ratios of formants appears inferior to simple formants until ratios representing all formant information ($F1$, $F2$, and $F3$) and $F0$ is utilized. At that point, accuracy is approximately equal to simple formant performance. The performances of x, y, z and x', y', z' are virtually identical since both variable sets represent the same four parameters. However, the x', y' combination yields somewhat higher accuracy than the simple log ratios because $F0$, $F1$, $F2$, and $F3$ are all still represented in various proportions by these two coordinates.

Table 2.6: Linear discriminant analyses of plurality identifications.

Variable(s)	% Correct	Average *APP*
$F1, F2$	80.8	0.671
$F1, F2, F3$	85.3	0.700
$z : log(F2/F1)$	37.6	0.275
$y : log(F1/SR)$, z	81.6	0.694
$x : log(F3/F2)$, z	48.6	0.365
x, y, z	85.6	0.726
x', y'	61.5	0.496
x', y', z'	85.7	0.726

If the utilization of $F3$ in vowel perception is predominantly that of a retroflexion detector, then the addition of $F3$ to $F1$ and $F2$ alone in accounting for the data should be minimal when /ER/ identifications are removed from the data set. The results of linear discriminant analyses for such a data set are shown in Table 2.7. Note that not only are all indices higher, but the addition of $F3$ to $F1$ and $F2$ alone increases the percent correctly classified by only 0.4%. This suggests that indeed $F3$ may primarily be used by subjects to mediate the percept of retroflexion and that $F1$ and $F2$ are the primary attributes used to mediate non-retroflex vowels.

Besides providing an index of the relative group membership strength, reflected in the average *APP*, linear discrimination analysis can provide the *a posteriori* probability of group membership for individual tokens. Assmann, Nearey, and Hogan (1982) attempted to predict vowel identification responses by examining the correlations of *APP* scores for individual tokens from linear discriminant analyses with the identification rates of 100 natural isolated vowels and gated portions of the same vowels. They found correlations, using the Spearman rank order correlation statistic, ranging from $R^2 =$ 0.007 to 0.490, depending on the analysis variables and data set manipulations under test. They cited four reasons for considering this degree of correspondence to be noteworthy: 1) the *APP* scores were based on a small vowel sample compared to the listeners' presumably larger experience

Table 2.7: Linear discriminant analyses of plurality identifications (no /ER/).

Variable(s)	% Correct	Average *APP*
$F1, F2$	87.0	0.729
$F1, F2, F3$	87.4	0.734
x', y'	61.6	0.494
x', y', z'	87.1	0.755

base; 2) the acoustic variables used in the analyses may not have been those used by the perceptual system; 3) errors in measurement could not be ruled out; and, 4) context effects may have been present which can influence vowel identification.

If a sufficiently high correlation were found between *APP* scores for individual tokens and identification plurality rates for the current data, the *APP* scores could serve as a powerful predictive tool for establishing a saliency gradient, as was discussed in sections 2.3.6 and 2.3.7, for the vowel zones in *APS*. However, a correlation higher than those found by Assmann and colleagues would be required to serve any useful purpose.

The *APP* score for the classification of each token (except rejected points) was first obtained from the linear discriminant analysis yielding the highest index values (x', y', z'). A Spearman rank order correlation was then computed between the plurality frequencies expressed as a proportion of the total possible number of responses (16) and the corresponding *APP* scores of all tokens. A moderately low correlation (R^2= 0.381) was found after correction for ties. Although this low correlation suggests that the *APP* scores are not adequate for use in a predictive model for identification rates, it does agree well with the correlation (R^2= 0.373) found by Assmann et al. (1982) for identification rates of isolated vowels presented in a mixed speaker condition and *APP* scores generated using natural log-transformed values of $F1$, $F2$, and $F3$.

In summary, linear discriminant analysis is a statistical classification tool capable of providing relative indices of resolution and group membership strength based on *a posteriori* knowledge of the distributions of the variables under test. Analyses of this type on the current data suggest that a relatively high level of correct classification for vowel identifications

can be achieved utilizing $F1$ and $F2$ as variables and that the addition of $F3$ as a variable provides only minimal improvement except in the classification of the retroflex /ER/. Correct classification utilizing log ratios of formants represented by x, y, and z, or transforms of these variables represented by x', y', and z' is approximately equivalent to classification utilizing the first three formants. While the *a posteriori* probabilities for individual tokens have been suggested as possible predictors of identification rates, correlation between these probabilities and pluralities for the present data are low, suggesting that such a strategy is unsuitable as a predictive model.

2.3.10 Agreement by z' plane

As has been mentioned previously, movement along z' in the *APS* primarily reflects a change in $F3$ when *SR* is fixed, such that $F3$ increases as z' increases. Although the data presented thus far suggests that $F3$ is not influential in the perception of non-retrofex vowels, it is of interest to determine whether or not $F3$ plays at least some role beyond the retroflex/non-retroflex distinction in determining the perceptual salience of vowels.

To investigate this issue, the minimum, maximum, and average z' coordinate values were found for all vowel measurements in each of ten vowel categories from the CID natural speech database and from Peterson and Barney (1952). These values, along with the z' range expressed in log units, are shown in Table 2.8 for all vowel categories and the averages of these values across all non-retroflex vowel categories. The data from Table 2.8 indicates that for natural, non-retroflex vowels the highest minimum and maximum values of z' are found for the high front vowel /IY/ and gradually decrease as the tongue position moves down, back, and up again, with the lowest z' values found for the high back vowel /UW/. The average z' values however do not reflect the gradual slope of the minimum and maximum values, staying relatively constant around $z' = 0.70$ except for the extreme high vowels. The retroflex vowel /ER/ has the lowest minimum, maximum, and average z' values of all the vowel categories considered. The range of z' values appears to be somewhat smaller for front vowels and larger for back vowels, with an average z' range of .141 log units.

The average percentages of agreement by z' plane for all vowel categories were calculated from pairwise comparisons of all subject response sets and are shown in Table 2.9. The value

of $F3$ corresponding to each z' plane is shown in parentheses. To normalize across the

Table 2.8: Averaged values of minimum, maximum, and average z' for CID Natural Speech Database and results from Peterson and Barney (1952).

Vowel Category	z' min	z' max	$\overline{x}$ z'	z' range
IY	.679	.796	.729	.117
IH	.657	.790	.708	.133
EH	.655	.767	.705	.112
AE	.645	.784	.700	.139
AA	.622	.777	.703	.155
AH	.633	.774	.699	.141
AO	.620	.783	.699	.163
UH	.622	.773	.684	.151
UW	.612	.768	.671	.156
ER	.531	.727	.596	.196
$\overline{x}$ (No ER)	.647	.764	.700	.141

differences in response frequency of the vowel categories, the agreement percentages for individual vowel categories were calculated as the averaged sum of the agreements between each subject response set pair for that category each divided by the total number of different tokens identified as that category across the response set pair. If the assumption is made that synthetic tokens which reflect formant patterns found in natural speech are more perceptually salient than those tokens which do not, and if subjects' agreement on the identification of a token in some way reflects the salience of the token's vowel quality, we might anticipate higher percentages of agreement for tokens that follow the z' patterns found in natural speech. For the present data, we might expect to find higher percentages of agreement for each vowel category between the minimum and maximum z' values for natural vowels in that category and the highest percentage of agreement for tokens located nearest the z' average. Note that at the most general level, the total average percentages of agreement in Table 2.9 (bottom row) do not differ substantially across z' planes, suggesting that subjects agree with one another about 64% of the time no matter what the value of $F3$. However, comparison of Tables 2.8 and 2.9 at the category level indicates that synthetic

Table 2.9: Average percentages of pair-wise agreements for all subject response sets by z' range.

Vowel Category	z' plane (F3 in Hz)						
	0.50 (1137)	0.55 (1390)	0.60 (1697)	0.65 (2071)	0.70 (2528)	0.75 (3086)	0.80 (3767)
IY	0.3	6.4	21.1	41.9	52.1	54.3	49.6
IH	0.0	0.5	5.0	12.9	22.7	26.5	18.4
EH	0.0	0.0	8.4	29.1	36.4	39.7	34.1
EY	0.3	0.6	2.8	22.1	38.0	13.2	7.9
AE	0.4	0.5	15.3	59.3	64.8	62.1	51.6
AA	45.8	66.7	62.9	60.6	63.4	65.7	60.3
AH	24.1	29.7	36.7	44.3	49.2	47.5	41.0
AO	47.0	49.2	49.2	49.7	49.7	50.6	49.4
OW	52.5	54.4	57.2	53.0	56.0	57.0	47.6
UH	10.6	18.0	26.8	38.5	36.3	31.3	24.1
UW	72.2	70.7	69.4	67.8	64.2	57.0	47.0
ER	40.6	47.7	35.5	3.7	0.0	0.0	0.0
$\overline{x}$ (all categories)	29.4	31.3	32.5	40.2	48.4	45.9	39.2
$\overline{x}$ (all tokens)	65.5	67.1	63.4	64.0	66.5	64.6	57.7

tokens falling within the approximate $F3$ range found for natural tokens are generally agreed upon to a higher extent than are tokens outside this range. This suggests that tokens having $F3$ values approximating those of natural vowels are more salient than those tokens that do not. While this appears to be true generally, in that the highest total average percentage of agreement for all categories occurs for the $z' = 0.70$ plane, some differences are present at the individual category level which are of interest.

At the individual vowel category level, the data in Table 2.9 appear to be in good agreement with the minimum, maximum, and average z' values found in Table 2.8 for the vowels /IY/, /AE/, /AH/, and /UH/. For the front vowels /IH/ and /EH/, the highest agreement percentages occur in the $z' = 0.75$ plane, higher than the z' average of the natural data for these vowels, and the ranges of better agreement along z' also appear to be shifted up. The mid and low back vowels /AA/ and /AO/ do not appear to reflect the natural data in that the agreements are relatively uniform across all z' planes and no distinct plane of greatest agreement is apparent. Although we have no natural data for the mid vowels /EY/ and /OW/, the mid front vowel /EY/ has its maximum agreement percentage at $z' = 0.70$, but good agreement is over only a very small z' range ($z' = 0.65$ to 0.70), while the mid back vowel /OW/ has more uniform agreements similar to /AA/ and /AO/. The high back vowel /UW/ exhibits a different agreement pattern across z' planes from all other vowels. While relatively good agreement for /UW/ is found for all z' planes, the highest agreement is at $z' = 0.50$, well below the z' average for this vowel category. Agreements for /UW/ then become progressively poorer with increasing values of z'. Agreements for the retroflex vowel /ER/ appear to generally agree with the natural data in range, although the z' plane of maximum agreement is slightly lower than the z' average for the natural data.

If $F3$ plays no role in determining the salience of a vowel beyond the retroflex/non-retroflex distinction, the good agreement between the natural data and the agreement percentages for the front vowel categories may still simply be due to the fact that the values of $F3$ for the planes where $z' = 0.65$ or greater are sufficiently high to accomodate the relatively high $F2$ values required for perception of these vowels. However, if this is the case, we might anticipate finding that all other vowel categories equally salient across z' ranges, since the values of $F2$ associated with their perception should be less than 1137 Hz,

the lowest value of $F3$ utilized in the experiment. While this could be true for the /AA/, /AO/, and /OW/ categories, it does not appear to hold for the other mid and back vowels. Thus, while the identification of a vowel token may be predominantly determined by the values of $F1$ and $F2$, $F3$ may play some role in the perceptual saliency of the token, if saliency and subject agreement are related.

In summary, there is general agreement between the ranges of z' for natural vowels and the ranges of z' for higher average identification agreements of synthetic vowels, although specific differences exist between. This suggests that changes in $F3$ have an effect on the saliency of tokens, even though the identifications for non-retroflex vowels can be well accounted for by $F1$ and $F2$ alone.

2.4 Comparisons of vowel classification schemes

A number of theories of vowel recognition based on acoustic attributes of the speech signal have been developed in the past in an attempt to normalize across inter-and intra-speaker differences and, in so doing, resolve the overlap often found between vowels of differing quality when they are grouped along various acoustic dimensions. These theories have sometimes been categorized into two groups determined by the type of information postulated as utilized by the listener in vowel perception (Ainsworth, 1975; Neary, 1989). Theories requiring *extrinsic* specification assume that information from a number of vowels of a single talker are utilized to establish a reference for perception. Theories of this kind would seem difficult to utilize for classification of some synthetic data since token construction may not closely follow parameters of natural speech or of any one talker. However, several hypotheses for the provision of extrinsic information from such data will be tested using one classification scheme based on extrinsic specification. The second group of theories assume that all information necessary for vowel recognition is *intrinsically* specified within the speech signal itself. This section will discuss and evaluate how well a number of these schemes are able to classify the perceptual data from Experiment I and the natural sppech data from Peterson and Barney (1952) as compared to the *SSB* target zones.

2.4.1 $F1$ x $F2$

An early approach to intrinsically specified vowel classification was utilized by Peterson and Barney in 1952. This approach consisted of plotting measurements from spectrograms of $F2$ against $F1$ for ten vowels from 76 speakers and hand-drawing ellipses around data points of like vowel quality. The plots utilized a frequency scale devised by Koenig (1949) which is linear to 1 kHz and logarithmic above. Although the number or percent of tokens correctly classified by this procedure was not reported in this study, Peterson and Barney do state that the ellipses enclosed approximately 90% of the data points in this manner. They go on to state that considerable overlap exists between /ER/ and /EH/, /ER/ and /UH/, /UH/ and /UW/, and /AA/ and /AO/, although the overlap between /ER/ and other vowels may be disambiguated by taking the low values of $F3$ found for /ER/ into consideration.

The outlines of the ten vowel ellipses (see Figure 2-14) from Figure 9 of the Peterson and Barney study were traced using a digital tablet and stored on a computer for use in a plotting program to generate the figures that follow. The plurality identifications for the 1674 synthetic tokens (rejected points not included) were then plotted, overlayed on the ellipses, in a $F2$-$F1$ space using the Koenig scale for the $F2$ axis. The results of these plots are shown in Figure 2-15 for all categories except /ER/ and Figures 2-16a-g for all tokens by z' plane.

Although there is considerable overlap, the data points seen in Figure 2-15 fall generally in the appropriate ellipses and reasonably good agreement is found between the /AE/, /AA/, and /UW/ data and their ellipses. The upper portions of the ellipses for the front vowels contain no data points, presumably due to the fact that these spaces originally enclosed the higher combinations of $F1$ and $F2$ for women and children in the Peterson and Barney study and such combinations were not within the constraints allowed for synthesis. The /EY/ and /OW/ identifications, categories not used in their study, generally share the ellipses of neighboring categories, /IH,EH/ and /AO,UW/ respectively, as might be expected. With $F3$ information not utilized, identifications from five non-retroflex vowel categories fall in the ellipse for /ER/.

Although no attempt was made to determine the percent correctly classified, classifica-

Figure 2-14: Vowel ellipses plotted in $F1$ x $F2$ space from Figure 8 of Peterson and Barney (1952).

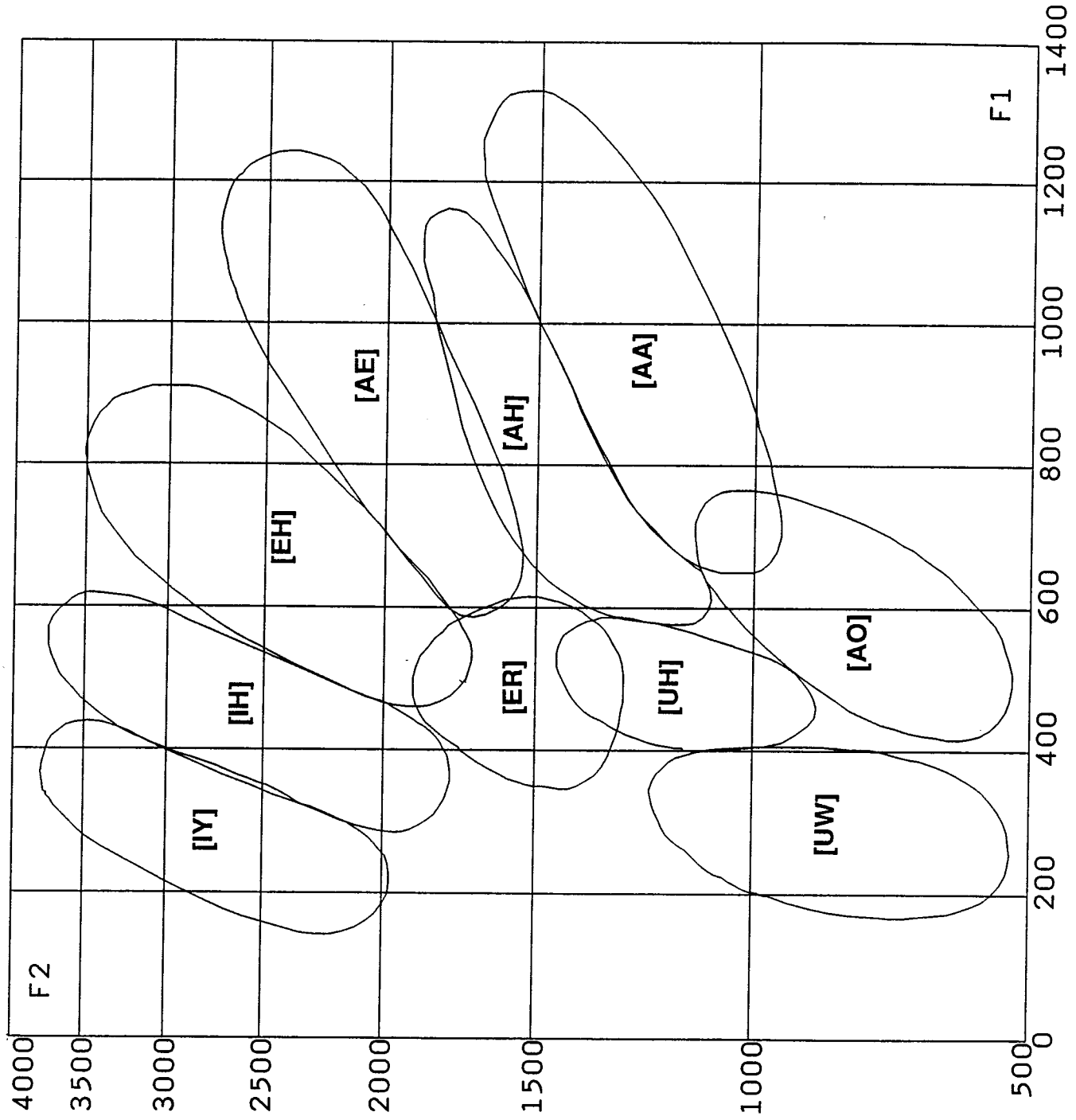


Figure 2-15: All plurality identifications with ellipses from Figure 2-14 in $F1$ x $F2$ space.

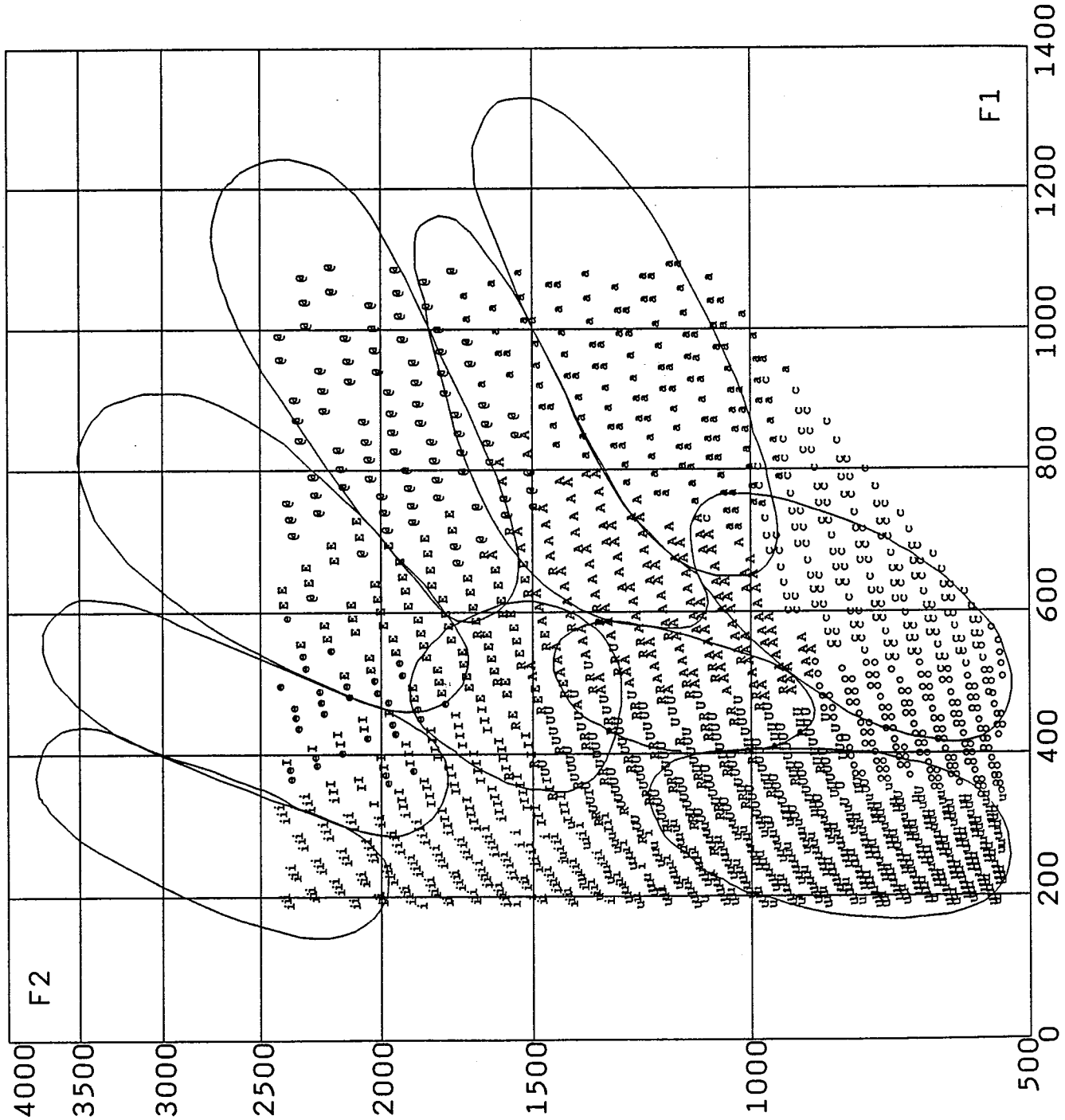


Figure 2-16: (a) Plurality identifications from the $z' = 0.80$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

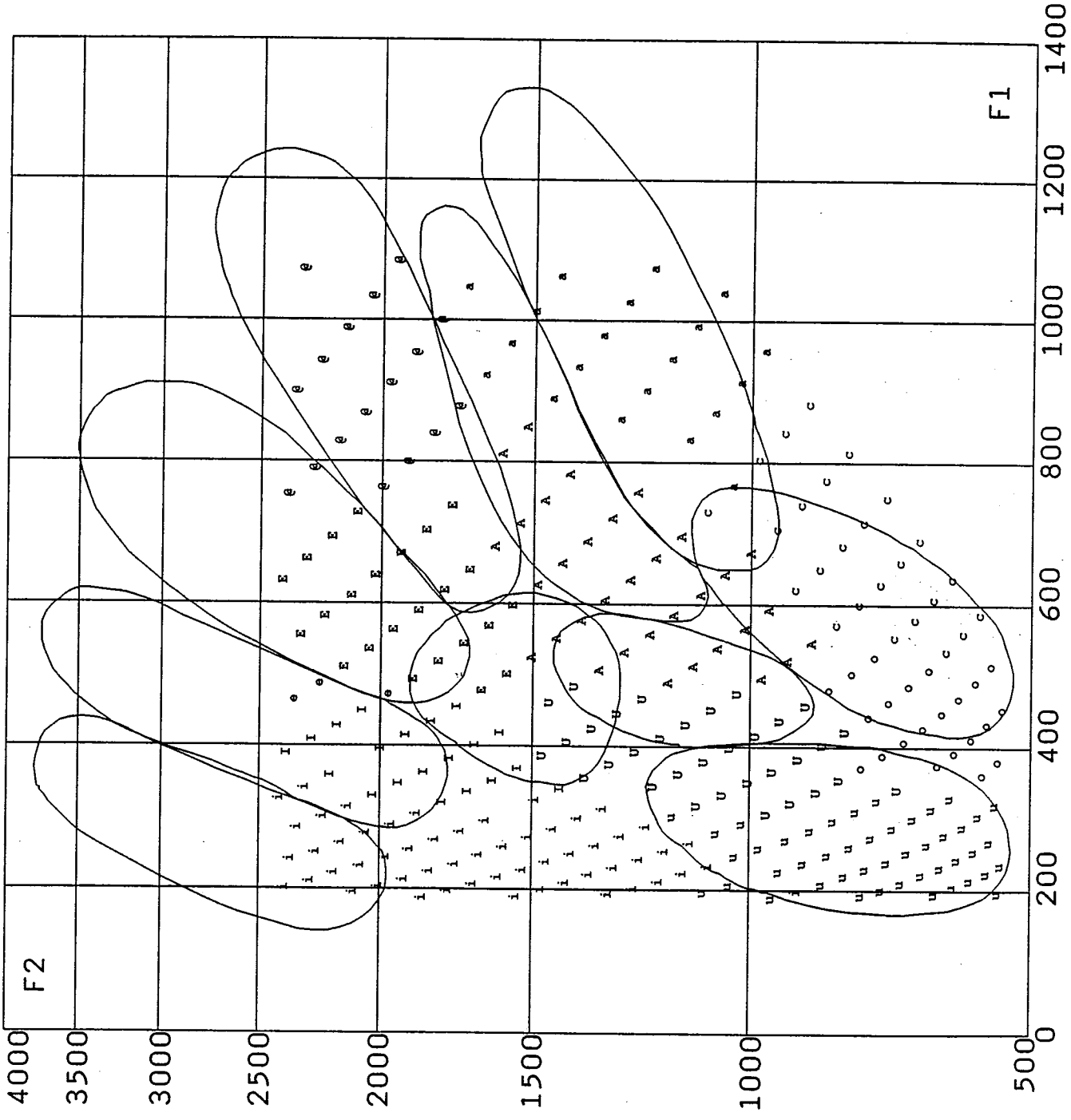


Figure 2-16: (b) Plurality identifications from the $z' = 0.75$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

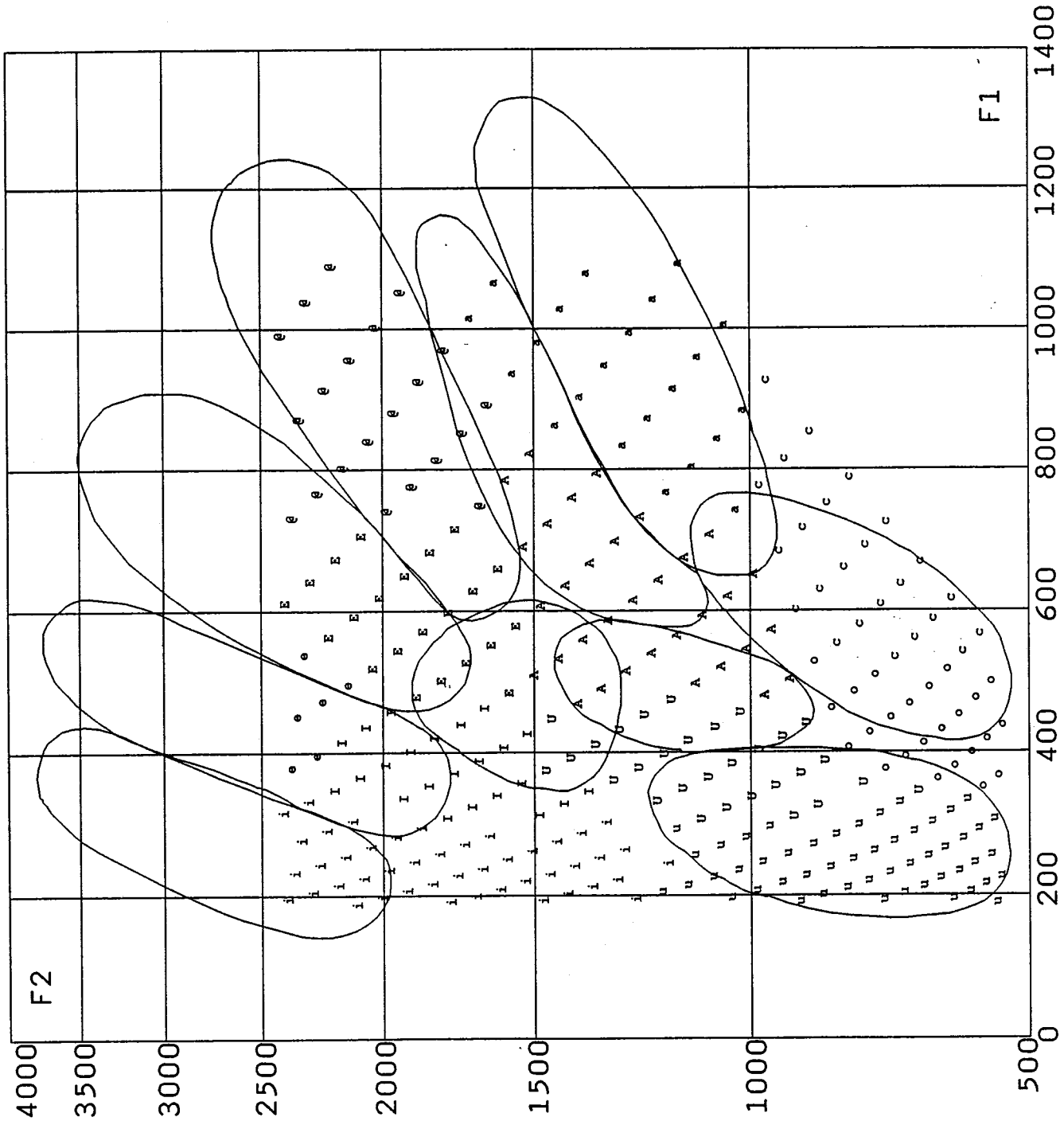


Figure 2-16: (c) Plurality identifications from the $z' = 0.70$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

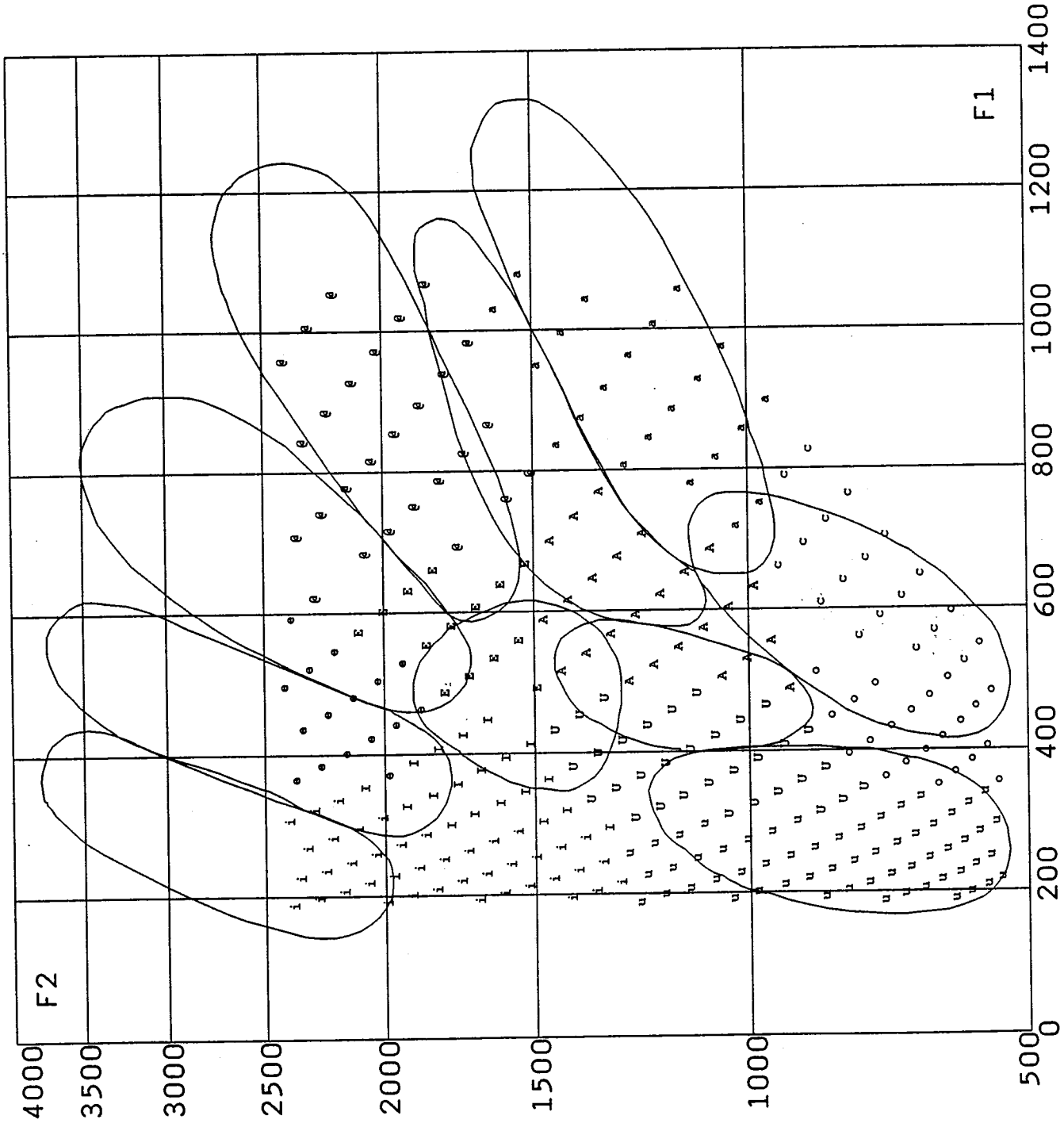


Figure 2-16: (d) Plurality identifications from the $z' = 0.65$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

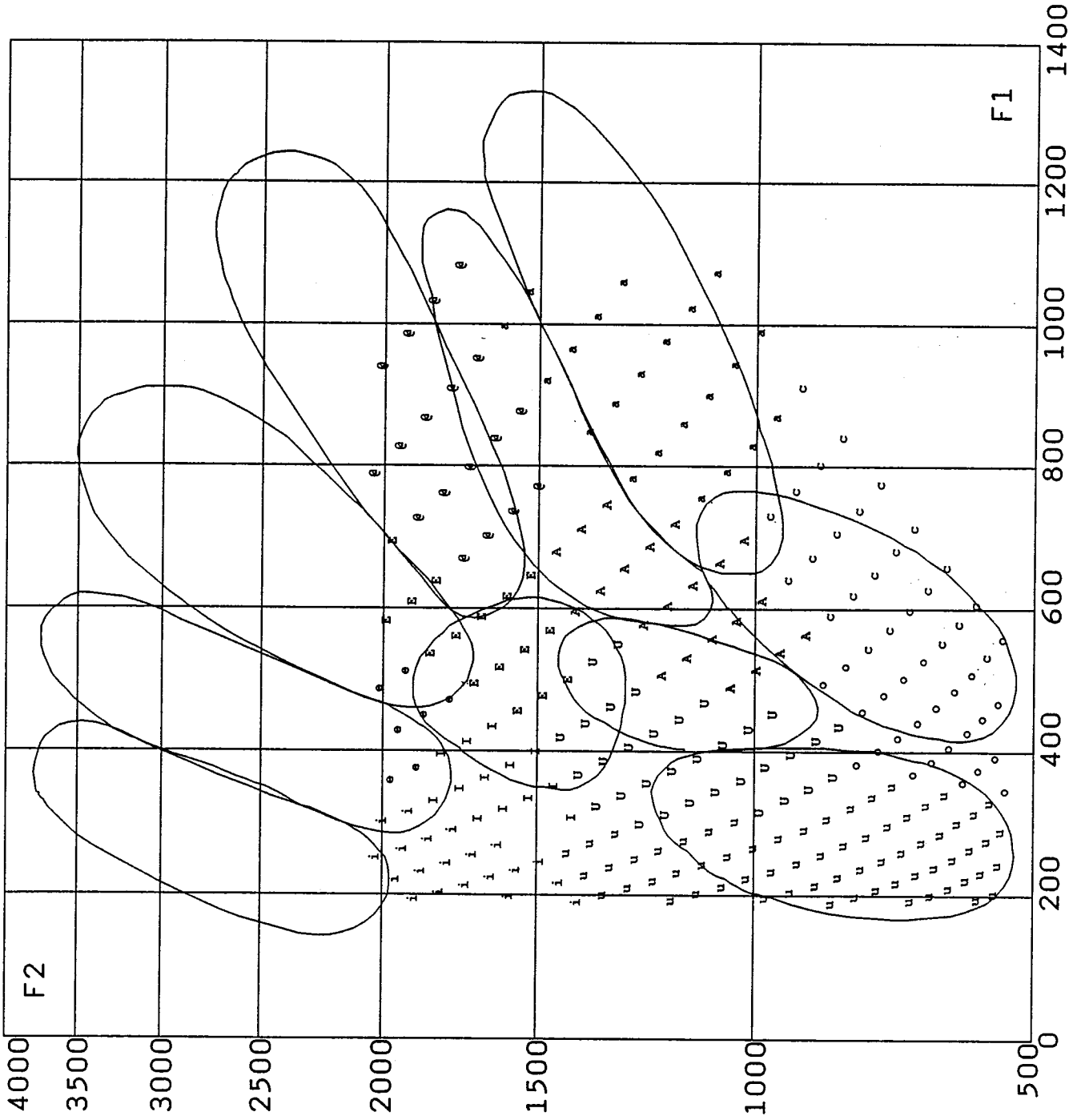


Figure 2-16: (e) Plurality identifications from the $z' = 0.60$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

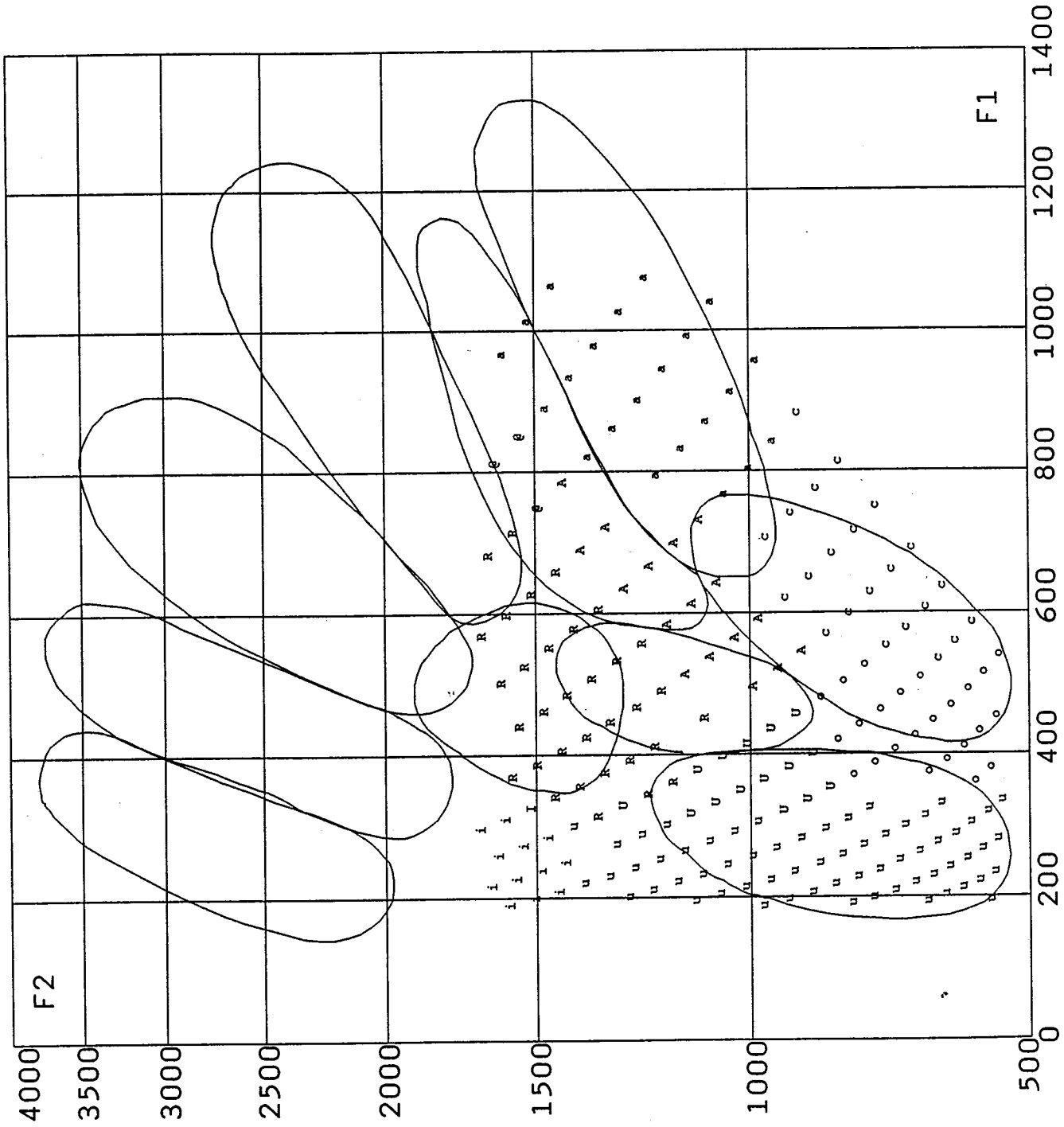


Figure 2-16: (f) Plurality identifications from the $z' = 0.55$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.

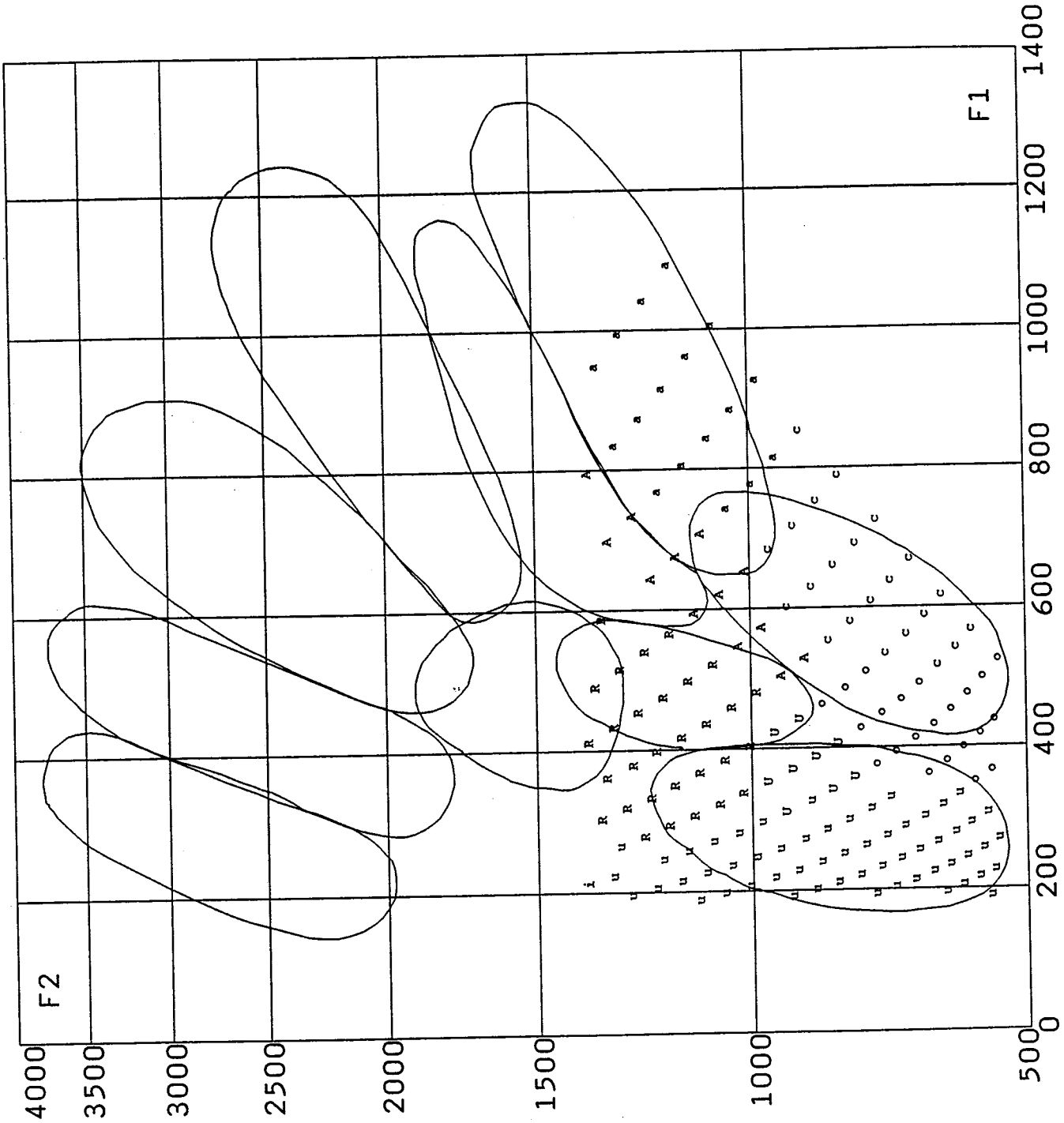
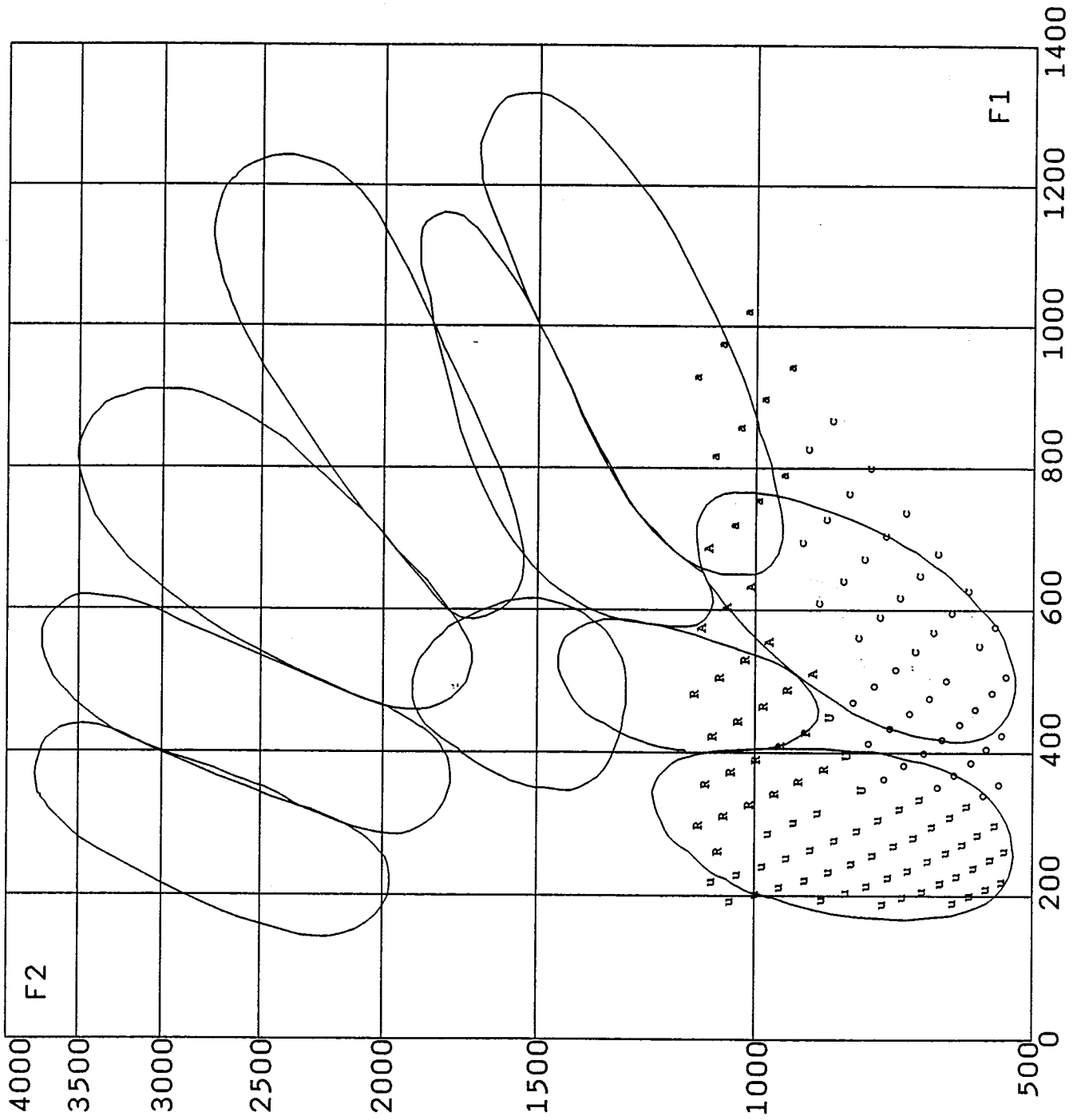


Figure 2-16: (g) Plurality identifications from the $z' = 0.50$ plane with ellipses from Figure 2-14 in $F1$ x $F2$ space.



tion accuracy does not appear to significantly improve for even the primary planes, seen in Figures 2-16d-f, where tokens should most closely approximate natural speech. Considerable scattering of the data into ellipses adjacent to the appropriate categories is found for virtually every vowel group. Although new ellipses could be drawn based on the synthetic data with perhaps some improvement in classification accuracy, no simple delineation of vowel categories using this method is apparent that would yield high classification accuracy.

2.4.2 Comparison of synthetic and natural speech-based target zones

The vowel classification accuracy using the natural speech-based (*NSB*) target zones and the synthetic speech-based (*SSB*) target zones will now be estimated and described for several different data sets. These classifications are made possible by a computer program which creates three-dimensional digital maps of the zones using *APS* coordinates and then references these maps for the locations of labeled vowel tokens. The mapping resolution employed was 0.001 log units for both the *NSB* and *SSB* zones. Data points falling on zone boundary lines are counted as belonging to that zone. Since the shapes of *SSB* zones change at 0.05 log unit intervals in the z' dimension, the way in which the classification program deals with data points lying between the specified intervals is of importance. The program does not interpolate the zone boundaries between specified z' planes, but rather, utilizes the specifications nearest the data point of consideration. Thus a data point falling less than 0.025 log units behind a specified z' plane would be classified with the zones mapped from the z' plane in front of it. The *NSB* map boundaries reflect those shown in Figure 1-9 from Chapter 1. Map areas between boundary lines or outside zones are considered unclaimed space.

The preliminary results of classification are shown in Table 2.10 for the *NSB* target zones and for the *SSB* target zones. The data set "ALL" contains all 27,600 responses from the 16 subject response sets. The "PLURALITY" data set contains the plurality identifications excluding the rejected points. The "AGREE" data set contains the 320 tokens having unamimous identification agreement. For each of these data sets, there is an additional set labeled "(primary)" which contains the same data limited to the primary planes. Additionally, the "CID" data set contains 599 points representing measurements

Table 2.10: Preliminary vowel classification using *NSB* and *SSB* target zones.

Data Set	N total	*NSB* zones # corr.	*NSB* zones % corr.	*SSB* zones # corr.	*SSB* zones % corr.
ALL	27600	5237	19.0	20424	74.0
ALL (primary)	14112	4165	29.5	10623	75.3
PLURALITY	1674	354	21.1	1674	100.0
PLURALITY (primary)	862	278	32.3	862	100.0
AGREE	320	76	23.8	320	100.0
AGREE (primary)	170	62	36.5	170	100.0
CID	599	597	99.7	481	80.3
P & B	1520	1378	90.7	960	63.2

from natural speech samples and data from the literature, and the "P&B" set, 1520 points representing the natural speech measurements made by Peterson and Barney (1952) of 10 vowels in [hVd] context spoken twice by 33 men, 28 women, and 15 children. All data points are expressed as x', y', and z' coordinates.

The classification accuracy for the *NSB* target zones reflected by the percent correct scores from column 4 of Table 2.10 is relatively poor for all synthetic speech data sets compared to the scores for the *SSB* zones in column 6. However, the classification accuracy regarding these sets is underestimated, due to two factors. The first factor is that certain identification categories (/OW,EY/) used in the synthetic data sets are not currently represented by *NSB* target zones, and thus could not possibly be correctly classified. The number of identifications in the /OW/ and /EY/ categories for each data set are shown in column 3 of Table 2.11. When these tokens are subtracted from the totals, the resulting percentages of correct classification increase somewhat. The second factor concerns the difference between the space occupied by the *NSB* target zones and the space utilized for synthesis. Column 4 from Table 2.11 indicates the number of tokens falling in unclaimed space, i.e. areas between or outside of target zones. Detailed examination of the classifica-

Table 2.11: Corrected vowel classification using *NSB* and *SSB* target zones.

Data Set	N total	*NSB* zones # EY,OW	# UCS	corrected N	% corr.	*SSB* zones # UCS	corrected N	% corr.
ALL	27600	2902	13321	11377	46.0	752	26848	76.1
ALL (primary)	14112	1610	5720	6682	61.4	288	13824	76.8
PLURALITY	1674	171	804	699	50.6	—	—	100.0
PLURALITY (primary)	862	100	343	419	66.3	—	—	100.0
AGREE	320	10	198	112	67.9	—	—	100.0
AGREE (primary)	170	10	95	65	95.4	—	—	100.0
CID	599	—	1	598	99.8	10	589	81.7
P & B	1520	—	62	1458	94.5	71	1449	66.3

tion data indicates that less than 10 tokens fall between narrow boundary lines, thus the majority of unclaimed tokens fall outside the *NSB* target zones. If the number of unclaimed tokens along with the /OW/ and /EY/ identifications are subtracted from the totals of each data set, the percentages of correct classification increase considerably, although over half the data in each synthetic data set has now been excluded from consideration. Note that overall, the classification accuracy of the *NSB* zones for the synthetic data increases with increasing data set identification agreement.

The number of tokens falling in unclaimed space are also used to reduce the total number of tokens under consideration for the *SSB* zones. However, now the majority of tokens in unclaimed space represent only locations between zones, not outside of zones. This could provide the advantage of classifying these tokens with multiple identifications, rather than just listing them as unclaimed, although this will not be attempted at this time. The percent correct classification by the *SSB* zones of all identifications collected from Experiment I after correction for unclaimed space is 76.1%. This percentage should represent the maximum possible correct for classifying this data set, since the *SSB* zones were constructed to classify the plurality identifications from this data correctly, and the

remaining 23.9% of the identifications must represent the minimum number of differences from the plurality identifications. This implies that the percentages correct for the "ALL" data sets using the *NSB* target zones may actually reflect an even higher accuracy than indicated.

The data sets limited to identifications in the primary planes were included to evaluate whether or not the identifications of tokens located in the z' planes of the *APS* most associated with natural speech are classified with higher accuracy than identifications from all z' planes. For the *SSB* zones, only a very slight increase in accuracy is found when all identifications are limited to the primary planes. This reflects the results discussed in Section 2.3.10, where the average pair-wise agreement by z' plane remains relatively constant. However, for the *NSB* zones, an average increase in classification accuracy of over 15% is found for each data category when the data set is limited to the primary planes. This suggests that the *NSB* zones are more accurate at classifying identifications of synthetic tokens located in *APS* areas shared by natural speech, although the difference in accuracy may predominantly reflect the inability of the *NSB* zones to capture differences in the z' dimension outside the range of natural speech.

The classification accuracy for the natural speech data sets (CID and P & B) is somewhat reversed from that of the synthetic data sets. The *NSB* target zones appear to be considerably more accurate at classifying natural speech than are the *SSB* zones. The higher accuracy for the *NSB* zones could be anticipated, given the consideration that these zones were constructed using these data sets as their basis. However, the classification accuracy of the *SSB* zones is lower than anticipated and requires at least speculative explanation. Since the *SSB* zones were developed utilizing tokens approximating a male voice, the possibility that the female and child data may be the source of the poor classification accuracy was investigated. Separate classifications were run for the tokens from men, women, and children of the data set from Peterson and Barney (1952) with the *SSB* zones. The results indicated that the male data was classified only slightly more accurately (68.5%) than the overall perecentage and that the data from women and children slightly less accurately, 65.2% and 63.2% respectively. These differences in classification accuracy cannot be considered accountable for the low overall accuracy.

To further investigate the lower than expected classification accuracy of the *SSB* zones for natural speech data, the errors in classification by the *SSB* zones were more closely examined. This examination yielded two major sources of errors. The first error source was the erroneous classification of tokens as /EY/ or /OW/. The /EY/ classification errors accounted for 81.6% and 69.3% of the errors made on /EH/ and /IH/ identifications respectively, and the /OW/ errors accounted for 60.9% and 34.2% of the errors made on /AO/ and /UH/ respectively. Overall, tokens classified as /EY/ or /OW/ accounted for 41.7% of the total errors. If these errors along with tokens classified in unclaimed space are subtracted from the total possible, classification accuracy increases to 77.1%. These results suggest that tokens located in at least portions of the zones for /EY/ and /OW/ may be identified as vowels in neighboring zones when /EY/ and /OW/ are not among the possible responses for identification. Remapping these areas without /EY/ and /OW/ among the possible identification responses may clarify this issue and increase the classification accuracy of the *SSB* zones.

A second source of error gives consideration to the possibility that the resolution of the zones is too coarse to accurately estimate the true boundaries between zones. That is, the step size utilized in Experiment I for sampling the *APS* is too large to adequately reflect boundary information. The misclassifications for all vowels except /ER/ are shown as their intended identifications by grouping along the nerest z' axis in Figures 2-17a-c along with the zones appropriate for that z' region. Visual inspection of the misclassified tokens in *APS* indicates that the majority of errors not located in the zones for /EY/ or /OW/ occur at or near boundaries between zones. Examination of the plurality frequencies and ratings sums (See Figures 2-10a-g and 2-11a-g) along these boundaries suggest that some of these boundaries are weak, reflected by ties or low plurality and rating values, and that these boundaries may shift given a higher resolution mapping of these areas. A more complete appraisal of the classification accuracy of natural speech data with zones based on synthetic speech identifications should then perhaps rest with further research investigating boundary areas between zones in finer detail. The next section will discuss a preliminary attempt toward this goal.

Figure 2-17: (a) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.75$ plane which were misclassified by the *SSB* target zones.

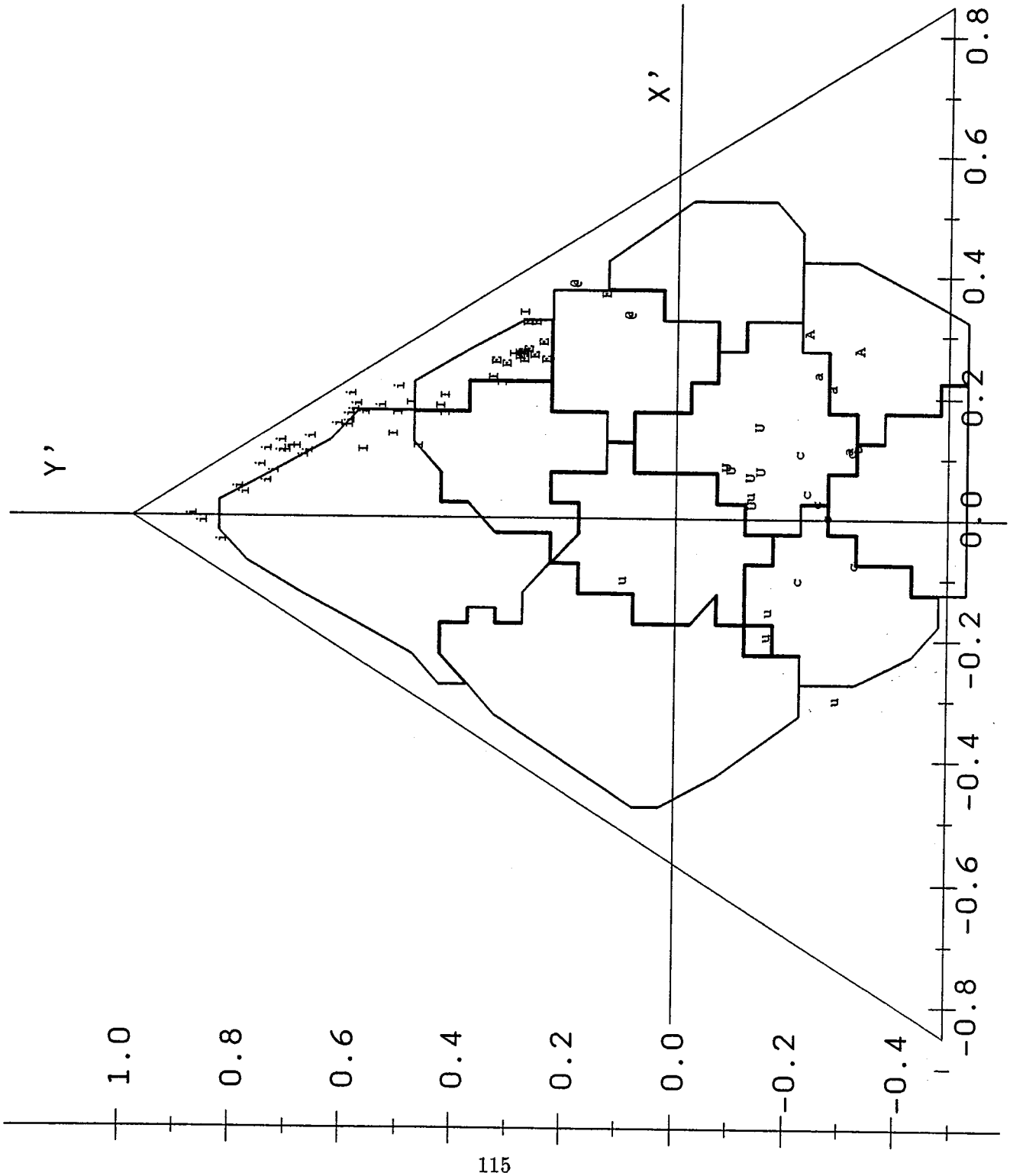


Figure 2-17: (b) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.70$ plane which were misclassified by the *SSB* target zones.

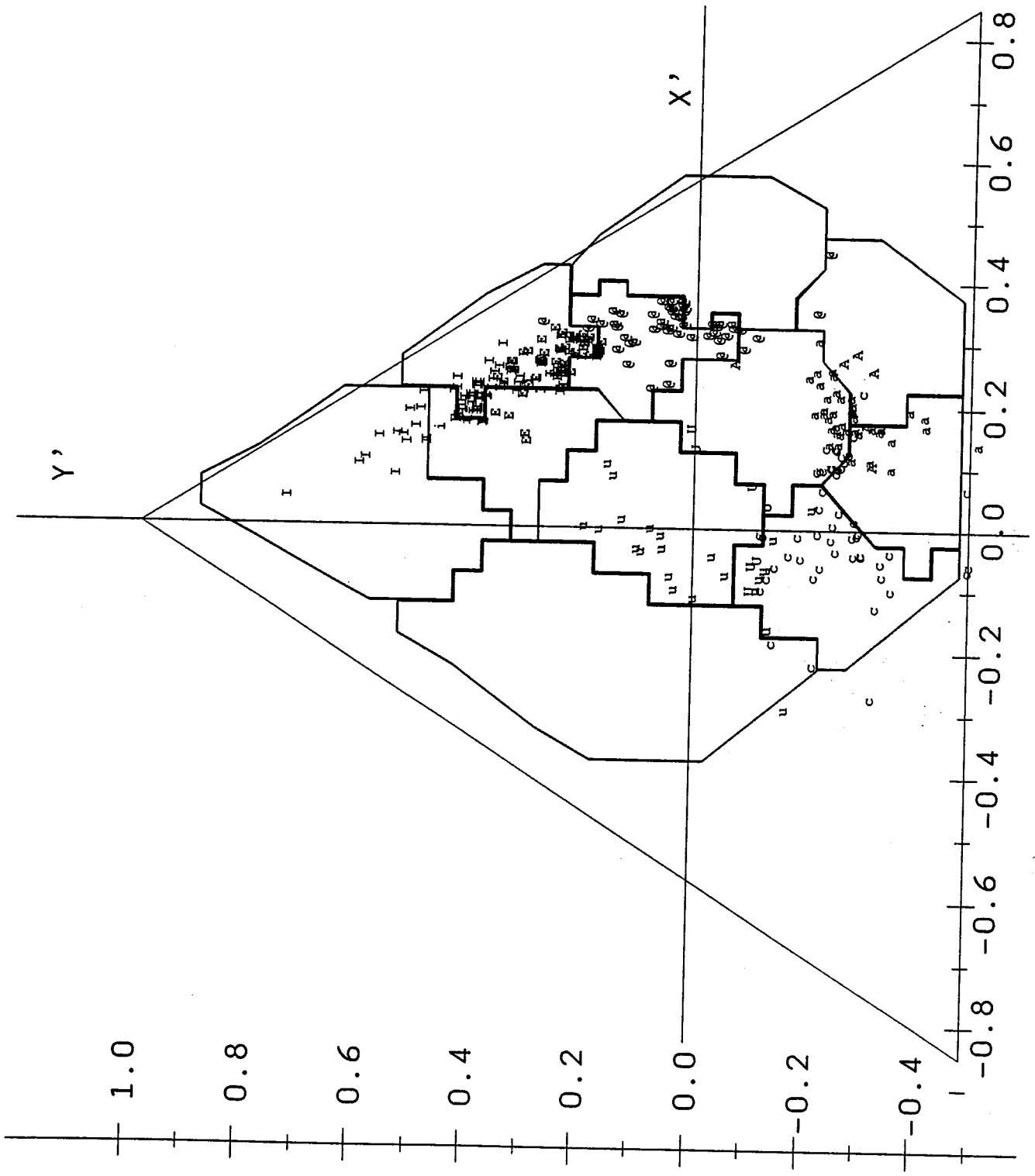
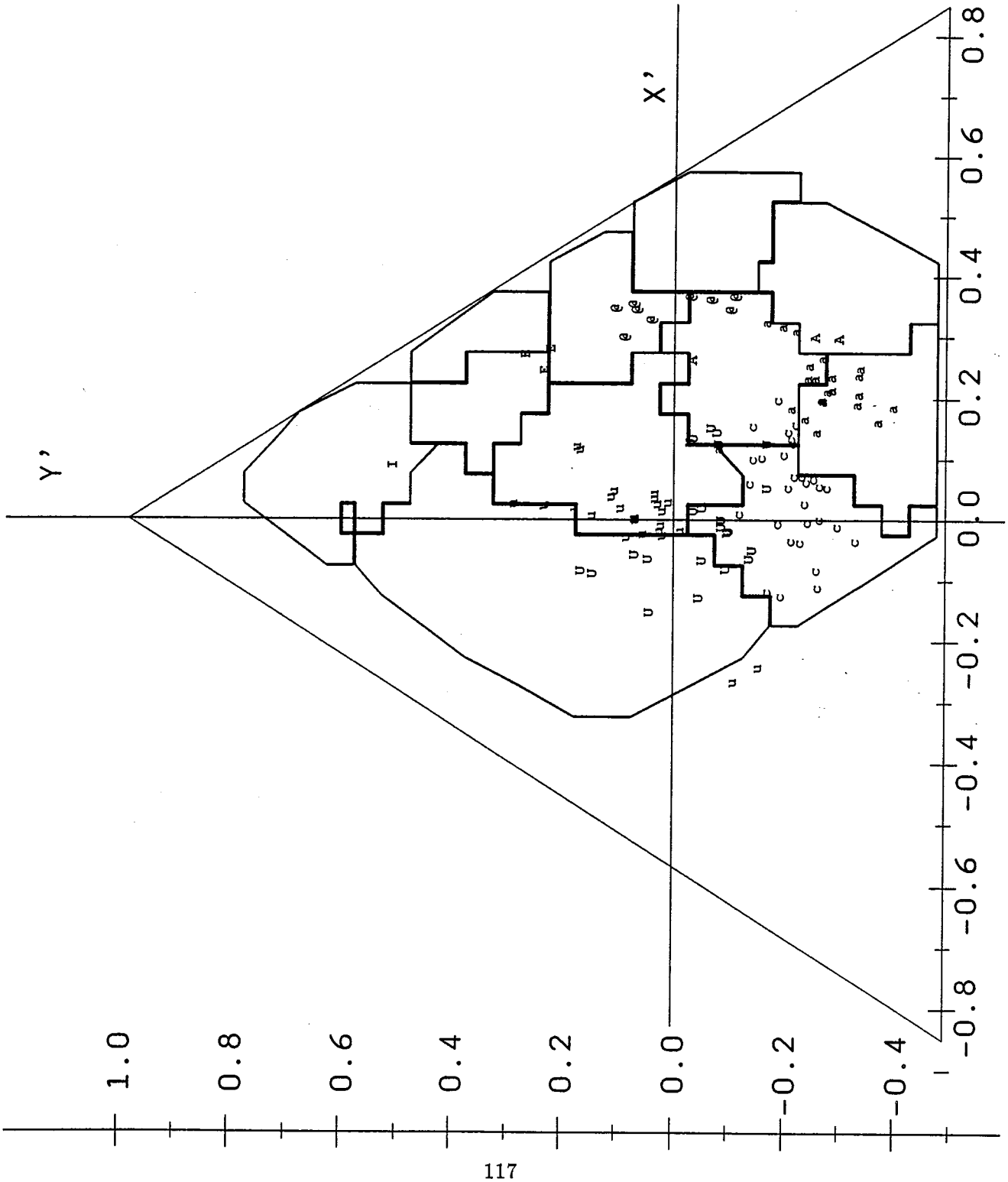


Figure 2-17: (c) Locations in APS $x'y'$ coordinates of vowel tokens from Peterson and Barney (1952) nearest the $z' = 0.65$ plane which were misclassified by the *SSB* target zones.



Classification using high-resolution target zones

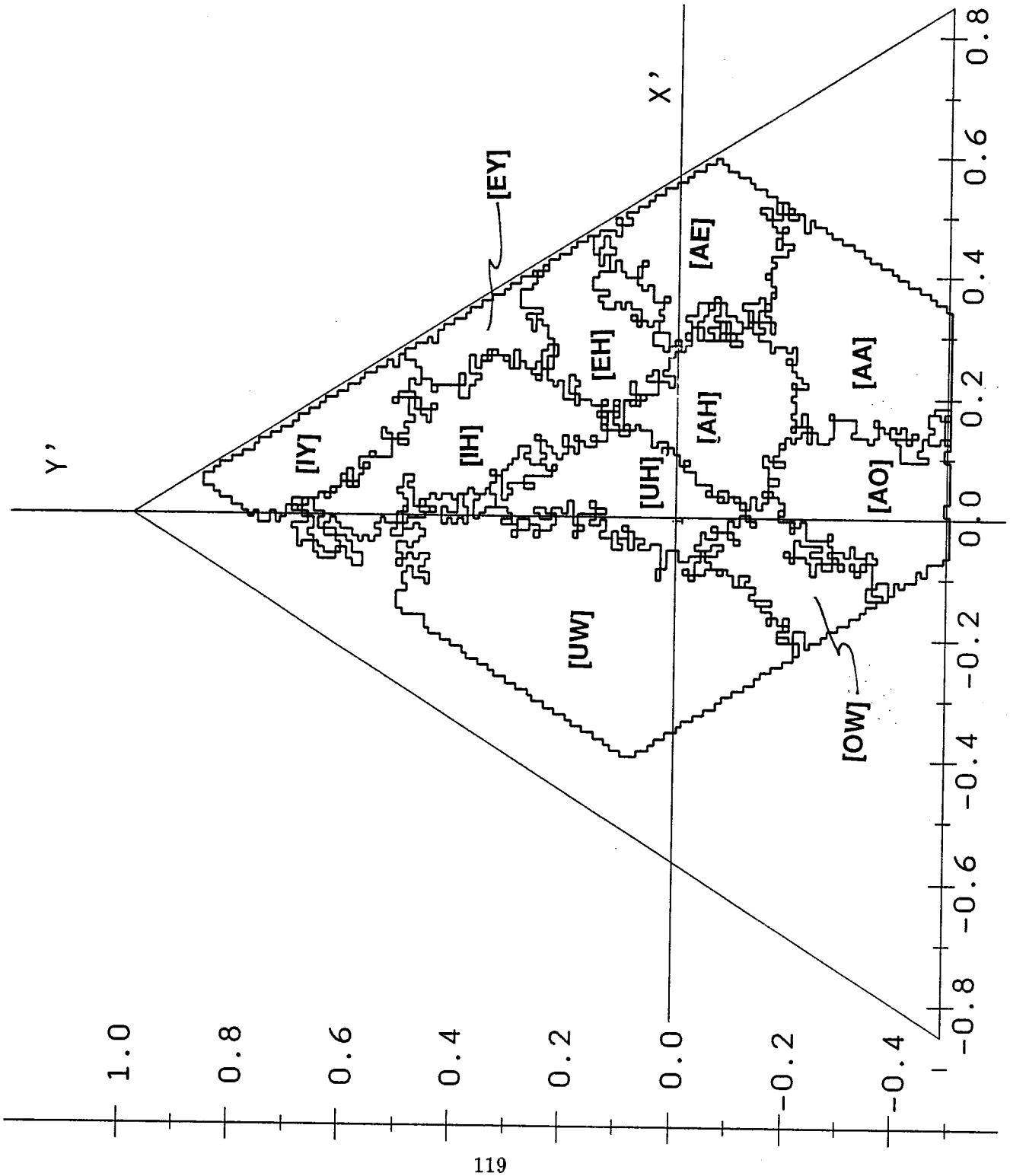
Miller and Hawks (1989) presented the results of a *APS* mapping experiment identical methodologically to the experiment presented in this paper with the exception of two differences. The synthesized tokens represented equidistant .01 log unit points in one z' plane ($z' = 0.70$) of the *APS* and were identified by only one subject highly trained in phonetics. This experiment yielded identifications to 7703 synthetic vowels for mapping this single z' plane at 5 times the resolution of the main experiment presented here.

Target zones, shown in Figure 2-18, based on these identifications were constructed automatically by a computer algorithm which encloses all points of like identification which are adjacent to one another. These zones were then used to classify the "PLURALITY" data limited to the $z' = 0.70$ plane and the "P&B" data limited to points located within the $z' = 0.70$ plane (± 0.49 log units). The results of classification with these zones (referred to as *HiR SSB*) along with similar results for the *NSB* and *SSB* zones are shown in Table 2.12. Note that classification accuracy for the "P&B" natural data set with the *HiR SSB* zones is over 12% higher than the lower resolution *SSB* zones. This result suggests that mapping at least boundary areas in *APS* at higher resolution may indeed increase classification accuracy of natural data. Additionally, the utilization of highly-trained subjects and equivalent response sets for identification of both synthetic and natural data may also aid in more accurate classification.

Table 2.12: Classification using *NSB*, *SSB*, and *HiR SSB* target zones.

Data Set	Zone Type	N total	# corr.	# UCS	% corr.
PLURALITY	*NSB*	296	105	-159	76.6
	SSB	296	296	—	100.0
	HiR SSB	296	224	-10	78.0
P & B	*NSB*	843	757	-30	93.1
	SSB	843	530	-18	64.2
	HiR SSB	843	629	-21	76.5

Figure 2-18: Location in APS $x'y'$ coordinates of computer-constructed high-resolution target zones based on data from Miller and Hawks, 1989.



2.4.3 Classification using bark differences

Syrdal and Gopal (1986) present an intrinsic vowel classification scheme which preserves articulatory feature information, and thus specifies vowel quality. These researchers found that, when $F0$, $F1$, $F2$, and $F3$ were transformed to critical band (bark) values, $F1-F0$, $F2-F1$, $F3-F2$ differences delineated the features high/non-high, compact/noncompact, and front/back respectively. The point of delineation for the binary feature distinction was 3 bark with differences less than 3 bark represented as a "+" for that feature and differences equal to or greater than 3 bark a "−". The resulting binary feature system for American English vowels[5] are shown in Table 2.13. Syrdal and Gopal state that the bark differences for $F1-F0$ corresponding to vowel height straddle this boundary for the vowels /AO/ and /ER/, and thus cannot be classified along this dimension, although both generally exceed the 3 bark criterion. The vowel /ER/ however can be solely classified on the basis of an additional spectral distance measure of $F4-F3$, should this information be available. The bark difference classification scheme was implemented on computer and applied to the synthetic data sets used in section 2.4.2. Values of $F0$, $F1$, $F2$, and $F3$ for the data were first transformed to bark values following the critical band scale approximation of Zwicker and Terhardt (1980). This approximation was modified with Traunmüller's (1981) low-frequency end correction for frequencies below 250 Hz. The $F1$–$F0$, $F2$–$F1$, and $F3$–$F2$ bark differences were then calculated and tested with the vowel feature classifications from Table 2.13. The results of these classification analyses are shown in Table 2.14. In comparing the percentages correct for the bark difference metric for the synthetic data sets with the results of the *NSB* zones in Table 2.11, we find that the classification accuracy of the bark difference metric is considerably higher for all synthetic data sets except the agreements limited to the primary planes, for which it is about comparable. The classification accuracy for the natural data set from Peterson and Barney (1952) is reasonably good, and improves further with the elimination of the /ER/ identifications to 89.0%. However, the classification accuracy of the bark difference metric is inflated due to limitations of the vowel features utilized in the classification. Inspection of Table 2.13 reveals that the feature specifications listed are not

[5]The features shown in parentheses were not included in Syrdal and Gopal (1986) and have been determined based on the features best fitting identifications in the synthetic data sets.

Table 2.13: Vowel feature system using bark-difference dimensions from Syrdal and Gopal (1986). Features in parentheses are based on best fit to synthetic data.

Vowel	$F1 - F0$ < 3 bark	$F2 - F1$ < 3 bark	$F3 - F2$ < 3 bark
IY	+	−	+
IH	+	−	+
EY	(+)	(−)	(+)
ER	(+)	−	+
EH	−	−	+
AE	−	−	+
AH	−	−	−
AA	−	+	−
AO	(−)	+	−
OW	(+)	(+)	(−)
UH	+	−	−
UW	+	−	−

Table 2.14: Vowel classification using Syrdal and Gopal (1986) classification scheme.

Data Set	# corr.	N total	% corr.
ALL	16857	27600	61.1
ALL (primary)	10276	14112	72.8
PLURALITY	1094	1674	65.4
PLURALITY (primary)	673	862	78.1
AGREE	255	320	79.7
AGREE (primary)	160	170	94.1
P & B	1291	1520	84.9

capable of distinguishing all twelve vowel categories, but rather, six vowel category groups. The ambiguities include the tense/lax pairs /IY-IH/, /EH-AE/, /AA-AO/, and /UH-UW/, as well as /ER/ and /EY/, which are indistinguishable from /IY/ and /IH/. Syrdal and Gopal (1986) state that this classification scheme cannot classify vowels properly without consideration of additional temporal parameters not captured by the bark difference metric. Since the synthetic tokens data classified here were constructed with identical duration parameters, this dimension cannot be utilized as a classification parameter. Thus, Table 2.14 actually reflects the classification accuracy of six categories of feature groupings and not 12 vowel categories. An accurate classification by vowel category utlizing this classification metric would yield correct classifications for only two categories /AH/ and /OW/, since all other categories are confusable. While such a metric demonstrates high sensitivity to vowel features, its ability to classify vowels by traditional categories is limited without the inclusion of additional information. This limitation suggests that this metric may not be appropriate for the classification of certain synthetic data.

2.4.4 Vowel classification utilizing extrinsic specification

As was mentioned previously, a number of vowel classification theories rely on information distributed across all vowels of a talker for normalizing or providing a reference framework for vowel classification. Often the information required is traditional acoustic parameters, i.e. formants, but some extrinsic theories also make use of formant bandwidths or formant amplitudes[6]. Collecting information from a reference series of vowels attributed to a single talker poses problems for classifying identifications for synthetic data as in Experiment I. This is because the motivation of the experiment was to uniformly map all the possible vowel space appropriate for a male talker and does not provide a set of reference vowels, per se. However, classification of the data sets used previously will be attempted with one such classification scheme utilizing two approaches for establishing a vowel reference framework for extrinsic specification. The classification scheme to be tested is from Neary (1977) and was selected because of its superior normalization performance in a comparative study by Disner, 1980. This scheme incorporates several different procedures, two of which will be

[6]For more detailed discussions of extrinsic classification schemes, see Disner, 1980 and Neary, 1989.

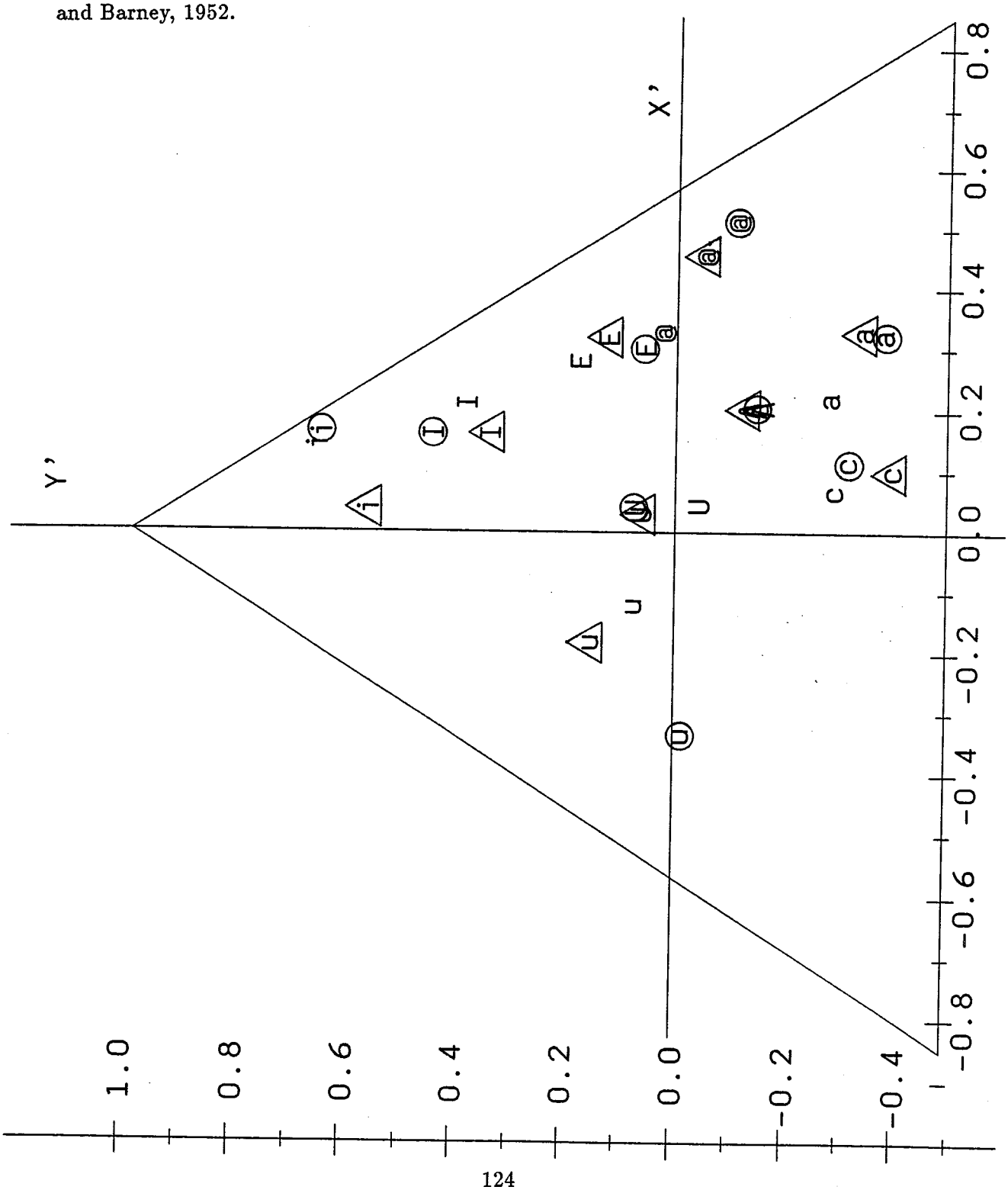
utilized here. Both procedures first transform formant values to their natural logarithms, notated $G1$, $G2$, and $G3$. The *CLIH* (constant log interval hypothesis) procedure then normalizes vowels by subtracting the average of the mean values of $G1$ and $G2$ for all vowels of a given talker from each $G1$ and $G2$. The *CLIH3* procedure is similar except that three independent parameters are established. Here the average value of $G1$ is subtracted from all values of $G1$, the average value of $G2$ is subtracted from all values of $G2$, and likewise for $G3$.

Two approaches for establishing a reference framework for extrinsic specification were utilized for classification. Both approaches assumed that all vowel tokens in Experiment I could be considered those of a single talker. The first approach hypothesized that the reference framework might be best represented by an exemplary token from each vowel category used in the mapping experiment. These exemplary tokens were determined by first selecting the token with the highest identification agreement for each vowel category from Experiment I. If several tokens were tied by this criterion, the token receiving the highest sum of ratings was selected. The averages of the natural log values of $F1$, $F2$, and $F3$ utilized in the synthesis specifications of these exemplary tokens served as the normalizing parameters for classification using this, the *EX* reference framework.

The second approach hypothesized that the reference framework might be best represented by the average location of all tokens identified as belonging to each vowel category used in the mapping experiment. The average x', y', and z' value was found for all tokens associated with each vowel category based on the plurality agreements. Formant values from these averages were calculated and the averages of their natural log values served as the normalizing parameters for this, the *AV* reference framwork.

The locations in *APS* of the token sets making up the reference frameworks for the two approaches are shown in Figure 2-19 along with the locations of the male averages from Peterson and Barney (1952). The vowel symbols in circles represent the first approach exemplary tokens, the symbols in triangles represent the second approach average tokens, and the plain symbols are the Peterson and Barney averages. Since the location of these points vary along the z' axis, manipulations were made to equalize these differences to a unitary z' plane. The locations of the Peterson and Barney points have been calculated

Figure 2-19: Locations in APS $x'y'$ coordinates of EXEMPLARY (circles) and AVERAGE (triangles) vowel reference frameworks used for extrinsic specification in Neary-type classification procedure along with locations (for comparison) of average male vowels from Peterson and Barney, 1952.



assuming a fixed $F0$ of 132 Hz, the value used for the synthetic tokens, instead of the actual average values. The values of $F3$ for all synthetic tokens and the Peterson and Barney data have been arbitrarily set to 2528 Hz, yielding a z' of .700 for all data points. This manipulation in effect allows the figure to reflect only $F1$ and $F2$ information in the relative distances between points.

Note in this figure that the tokens for the "point" or "corner" vowels, [IY,AE,AA,UW], representing extreme points of vowel articulation, are distanced furthest apart for the exemplary tokens than for the averaged tokens, and are closest together for the natural data tokens. Tokens representing vowels interior to the points of extreme articulation are generally more closely grouped across the three sets of tokens with a unitary location for the points representing the centralized vowel /AH/. These distance differences suggest that listeners may tend to separate the exemplary vowels at the extreme points of articulation, further apart than would be generally found in natural speech, in order to accomodate the vowel space being sampled. This follows the principle of maximum perceptual contrast, as demonstrated by Liljencrants and Lindblom (1972), whereby phonetic categories of vowel systems are distributed within the acoustic vowel space so as to maximize the distance between categories.

A similar expansion of vowel space can be seen in the data from Picheny, Durlach, and Braida (1986). In this study, the vowels [IY,AE,AA,UW] represented in a $F1$ by $F2$ space reach more extreme points of articulation when talkers are instructed to speak as clearly as possible, compared to the same vowels when spoken in a conversational manner.

The extrinsic normalizing parameters for classification of the natural data set from Peterson and Barney (1952) were established for each individual talker by calculating the average $G1$, $G2$, and $G3$ from the two vowel sets of each of the 76 talkers. The vowel category /ER/ was not included in the normalizing parameters for the synthetic or the natural data sets and was therefore also excluded as a category from the classification procedure. This exclusion reduced the total number of data points for consideration from 1674 to 1592 for the PLURALITY synthetic data set and from 1520 to 1368 for the natural data set from Peterson and Barney.

Vowel classification was by means of the linear discriminant analysis procedure described

in section 2.3.9. The results of the classification analyses for the two normalizing approaches for the PLURALITY synthetic data set previously described and the P & B data set are shown in Table 2.15. Note from the results that classification accuracy of the CLIH method

Table 2.15: Vowel classification using Neary (1977) classification scheme.

Data Set	Ref. Frame	N total	CLIH # corr.	CLIH % corr.	CLIH APP score	CLIH3 # corr.	CLIH3 % corr.	CLIH3 APP score
PLURALITY	*EX*	1592	1379	86.6	.7513	1389	87.2	.7559
PLURALITY	*AV*	1592	1380	86.7	.7513	1389	87.2	.7559
P & B	P&B	1368	1199	87.6	.8063	1230	89.9	.8483

is only slightly improved with the CLIH3 method. In addition, note that the results utilizing the *EX* and *AV* reference frameworks are virtually identical. The lack of differences between these two approaches may be explained by the fact that while the tokens used for normalizing parameters in each approach appear quite different in location, the averages of the formants for each set represent points which are separated by only 0.02 log units in the *APS*. Thus it appears that either set may be used to yield reasonably high classification accuracy. However, if the results of this classification scheme are compared to the results of linear discriminant analyses on the synthetic data using $F1, F2, F3$ and x', y', z' previously shown in Table 2.7, we find no improvement in classification accuracy. This lack of improvement suggests that the extrinsic reference frameworks do not adequately reflect the data, or that, as was speculated initially in this section, this type of classification scheme is not appropriate for certain synthetic speech experiments.

2.5 Summarization and Discussion of Experiment I

In summary, this experiment has demonstrated that perceptual target zones for the vowels of American English can be constructed which span an extensive range of possible vowel

sounds. Additionally, these target zones are abutting and non-overlapping and correctly classify over 99% of the 1725 synthetic vowel-like sounds used in the experiment, based on identifications representing the plurality of subject's responses.

Comparisons of subjects' identifications indicated that subjects agreed with one another on about an average of 63% of the tokens and totally agreed with one another on about 19%. It is difficult to evaluate these percentages of agreement, since no attempt was made to determine how many tokens representative of non-American English vowels, ambiguous vowel sounds, or non-vowel sounds were included which could greatly influence identification agreements. As a reference, however, results from past studies utilizing identifications of natural vowels spoken in isolation reflect higher, but highly variable subject agreements, with identification error rates ranging from 43% (Strange, Verbrugge, Shankweiler, and Edman, 1976) to 3% (Kahn, 1978). While this large range may reflect differences in the stimulus parameters and experimental methodology utilized, these studies generally utilized stimuli which were produced with the clear intention of representing only salient American English vowels. Such a statement cannot be made for the present experiment.

Subjects agreed with themselves on identifications of about 75% of the tokens, a significantly greater amount than that found for between-subjects agreements. In addition, the identifications for some tokens of individual subjects consistently and confidently deviated from the plurality identifications suggesting that perceptual boundaries between vowels may reliably vary among individual listeners. This finding poses a problem for the development of zones to generically represent vowel classification by all listeners and may require the addition of other speech parameters (i.e., some "top-down" processing) to disambiguate token classification at zone boundaries.

If the number of responses comprising a plurality identification, the plurality frequency, is considered a measure of vowel saliency, we find that target zones are graded along this dimension, with the highest frequencies (i.e, most salient tokens) located generally more central to the zones and progressively lower frequencies (i.e., less salient tokens), toward zone boundaries. While this result may not come as unexpected, given the results of numerous studies exploring vowel boundaries with single continua, it does provide a more complete picture of the saliency gradients associated with the total areas of vowel cate-

gories and suggests that such gradients could be of some benefit applied to current methods of speech identification. Furthermore, while confidence ratings for identifications may have the potential to provide additional information to saliency gradients, the results of this experiment indicate that the subjectiveness of such a scale must somehow be reduced before such information can be obtained.

Comparison of the new target zones based on synthetic speech with current estimates of similar zones based on natural speech suggest that, while both sets of zones exhibit considerable amounts of overlap between like categories, both also demonstrate rather poor classification accuracy of data other than that utilized in their construction. The inaccuracy found for the synthetic-speech-based zones in classifying natural speech data may be attributable to the coarse resolution utilized in the mapping procedure. This resolution precludes the estimation of precise boundaries between vowel categories which would be required for higher classification accuracy and, as was preliminarily demonstrated, zones mapped at higher resolution can be considerably more accurate. Additionally, duplicating the subtleties of natural speech with synthetically-generated speech to a degree sufficient to assure that all elements of natural speech relevant to perception are intact is still a problem which cannot be dismissed.

Several reasons may account for the inability of the natural-speech-based target zones to provide high classification accuracy of the synthetic data. The first of these reasons reflects the original arguments that motivated the present experiment, that is, the utilization of insufficient amounts of data for uniform mapping of the vowel space combined with the uncertainty of using data from sources varying in analysis, formant measurement, and identification methodology, make zones estimates calculated in this way subject to noise and constant modification with the addition of new data. To maximize classification of the natural data used in their construction and minimize overlap between adjacent zones, the boundaries for these zones in the $x'y'$ dimension have necessarily become quite intricate and complex. Much of this complexity is due to the inclusion of outlying data points, even though much of the enclosed zone space is unaccounted for with data. A viable question thus becomes whether or not these zones should be modified to reflect at least some of the synthetic results.

An additional problem in classification of the synthetic data with the natural-speech-based target zones is their current inability to capture changes in zone boundaries related to the z' axis. Although this is a solvable problem, the zones are not currently able to capture shifts along the z' axis of constant values of $F1$ and $F2$ which tend to result in like vowel identifications, nor the more subtle shifts in zone boundaries which can result with changes in $F3$. A more detailed estimation of these zones in the z' dimension may not only descrease the complexity of boundaries as viewed in $x'y'$ planes, but also improve classification of the synthetic data.

The last reason to be discussed relevant to classification inaccuracies is a shared problem to both sets of target zones. This reason is based in the difficulties arising from the mismatch in the number of zones or vowel categories to be represented. While information concerning the locations and perceptual saliency of zones for [OW] and [EY] are of interest, these categories are not represented with natural-speech-based zones and have rarely been used for perceptual studies of vowel classification for American English in the past. Had the [OW] and [EY] categories been eliminated from the response set in the present experiment or included as zones for natural speech, we could anticipate considerably less confusion between these categories and their neighbors resulting in higher classification agreement.

Statistical classification procedures suggest that the plurality identifications of non-retroflex vowels in this experiment can be well accounted for by the frequencies of $F1$ and $F2$ and that the addition of $F3$ does not increase the relative classification accuracy unless the retroflex vowel category [ER] is included. While this finding suggests that $F3$ may perceptually function as a 'retroflexion detector' only, additional analyses of trends in subjects' identification agreements relative to $F3$ suggest that $F3$ does influence the saliency of vowels. In general, tokens with values of $F3$ in the range found for natural vowels are agreed upon by subjects to a greater extent than are tokens with values of $F3$ outside this range. Despite this probable influence on vowel saliency, $F3$ does not appear to greatly affect the identifications of non-retroflex vowels in American English. This finding casts some shadow of doubt on the necessity of representing $F3$, and in the case of the *APS*, a third dimension, in vowel classification. However, there remain several reasons for retaining $F3$ representation, despite its questionable utility for American English.

The first of these reasons is that, although boundary shifts in vowel categories due to changes in $F3$ are small, they do exist and thus may be best captured in multi-dimensional representations of the vowel categories. Small shifts in the perceptual boundaries between vowel categories related to changes in $F3$ have been reported by Fujisaki and Kawashima (1968), Holmes (1986), and Neary (1989). Additionally, it can be demonstrated that boundary shifts related to changes in $F3$ also occurs in the present experiment. If the locations of all tokens identified as non-retroflex vowels from the present experiment are recalculated to reflect a unitary value for $F3$, a single set of two-dimensional zones for one z' plane can be utilized for their classification. Errors found in this classification should reflect in a general sense any boundary shifts induced by changes in $F3$, since the multi-dimensional zones constructed in the present experiment can correctly classify 100% of these identifications when $F3$ varies. The determination of which z' plane of zones may best represent all tokens should not be extremely critical if the assumption that boundaries in terms of $F1$ and $F2$ values do not shift with changes in $F3$, although intuitively the z' plane most representative of $F3$ values found in natural speech seems implied. Since the average z' value for non-retroflex vowels in the natural speech data considered in Table 2.8 is 0.70, the zones associated from that plane will be utilized for the classification. The value of $F3$ for all tokens (excluding rejected tokens) identified by the plurality of subjects as non-retroflex vowels was set to 2528 Hz, forcing their z' location to 0.70, and values for their locations in x' and y' recalculated. Classification of these tokens with the synthetic-speech-based zones for the $z' = 0.70$ plane was 88.3%. This result suggests that identifications of almost 12% of tokens may have been perceptually influenced by the value of $F3$ and that, while multi-dimensional zones can easily take such influences into account, two-dimensional zones based on values of $F1$ and $F2$ may have difficulty accomodating perceptual shifts influenced by $F3$.

An additional reason for $F3$ representation takes into consideration the more global potential for a third dimension. A metric reflecting $F3$ information is able to provide the potential for establishing target zones, and therefore classification, of languages other than American English. In particular, classification of languages which necessitiate perceptual differentiation of the rounded/unrounded distinction for vowels may require $F3$ representa-

tion. The need for representing $F3$ information to increase vowel classification over $F1$ and $F2$ information alone has been described by Fant (1973) for Swedish and Pols, Tromp, and Plomp (1973) for Dutch. The necessity of $F3$ information, as reflected in the z' dimension of the *APS*, for the disambiguation of rounded and unrounded vowels in German has been demonstrated in Jongman, Fourakis, and Sereno (1989).

An additional consideration for representing $F3$ information in vowel classification is its potential as a normalizing factor for differences between talkers. Generally, the variance found in $F3$ for vowels (rhotacized vowels excepted) of an individual talker is quite small compared to the variance found across talkers (Fujisaki and Kawashima, 1968). Furthermore, $F0$ and $F3$ are usually highly correlated for an individual talker, reflecting the generally close relation between the sizes of the glottis and vocal tract. These facts have been employed in a number of studies investigating how changes in $F0$ and $F3$ may affect vowel perception and how information from these two parameters may be used in across-talker normalization strategies for vowel classification (See Neary (1989) for a discussion). While results from these studies are often conflicting, a consistent finding is that perceptual boundary shifts between vowel categories are greater with concomitant changes in $F0$ and $F3$ than with either alone (Fujisaki and Kawashima, 1968; Neary, 1989). However, the issue of how to best utilize both $F0$ and $F3$ in a talker normalization strategy remains largely unresolved. The range of $F3$ utilized in the present experiment is considerably greater than is likely to be found associated with an individual male $F0$ in natural speech, thus perhaps creating an environment too unnatural for evaluating target zones intended for natural speech. The talker-normalization parameter utilized in the *APS*, *SR*, is currently influenced predominatly by $F0$ in its calculation, although through its defined intention to represent a talker's vocal characteristics, could represent additional parameters, like $F3$, should normalization by these means prove superior.

Another relevant issue concerning the role of $F3$ in vowel perception is rhotacization, that is, the auditory property of "r-coloring" of vowels. While the vowel [ER] may be considered completely rhotacized, as in the word "bird", partially rhotacized, or r-colored, vowels also occur in American English, as in the words "beard, bared, bored." As was mentioned previously in Section 2.3.1, subjects had difficulty labelling tokens which were

neither clearly non-retroflex monophthongs nor [ER], implying that these tokens may have been representative of r-colored vowels. However, the partial rhotacization of vowels is generally considered a dynamic process, whereby the intitial vowel sound is non-rhotacized and becomes increasingly more rhotacized over time (Ladefoged, 1982, p. 207). Since there was no formant movement in the present stimuli, the tokens in question do not fit the general description of r-colored vowels, but rather, may represent steady-state versions of formant patterns associated with single points in time along the dynamic of the partial-rhotacization process. While it is not likely that such sounds would be found in natural speech, the acoustic and perceptual aspects of r-colored vowels in American English has not been extensively investigated. The *APS* can provide an excellent framework in which to base further investigation.

Several schemes for natural speech vowel classification were compared and contrasted with the synthetic-speech-based target zone approach in terms of classifying the plurality-based identifications from this experiment. Each of these other schemes had difficulties with this task due to either inherent flaws in the classification approach itself or theoretical constraints to natural speech which made them potentially inappropriate for classification of some synthetic speech. The $F1$ x $F2$ ellipses from Peterson and Barney (1952) are flawed in that identifications related to changes in $F3$ cannot be reflected. Previous discussion has related the importance of $F3$ representation in vowel classification. The classification scheme from Syrdal and Gopal (1986) based on patterns of binary vowel features related to bark-transformed formant differences is flawed in that additional speech parameters like vowel duration are required to disambiguate vowel categories. While such parameters may prove capable of aiding the classification of natural speech with this approach, these parameters are not available for the synthetic tokens utilized here. The last classification scheme evaluated, from Neary (1977), requires the extrinsic specification of information about the distribution of vowels for an individual talker to provide a reference framework for classification. Although across several approaches to establishing such a framework with the synthetic identifications were attempted, classification results utilizing these frameworks were no better than for statistical classification with simple formant values. These results suggest that such a classification scheme may not be appropriate for use with synthetic

speech experiments where tokens potentially do not represent those of an individual talker.

Chapter 3

Experiment II: Estimation of difference limen for distance (d) in the APS vowel space

3.1 Introduction

With the general areas of *PTZ*s for synthetic vowels in the *APS* established in Experiment I, a more detailed investigation of the locations of the boundaries between PTZs is required. However, before such a task can be undertaken, an additional question of some importance must be addressed. While we can surmise from Experiment I that .05 log unit resolution is too coarse for accurate boundary estimation since boundaries estimated with .01 log unit resolution from a phonetically trained subject increase correct classification of natural vowels substantially (Hawks and Miller, 1989), mapping the entire vowel space utilized in Experiment I at this resolution would increase the number of tokens to be judged by an order of magnitude. For example, at the .05 log unit resolution employed in Experiment I, 304 tokens were evluated in the $z' = .70$ plane. Mapping this same plane at .01 log unit resolution requires 7703 tokens. In addition, such a mapping may potentially reflect considerable amounts of redundant information should neighboring tokens be perceptually identical. Thus the question of what level of resolution is required for sufficiently defining

PTZs and their boundaries in the *APS* requires attention. One method of answering this question is to determine the difference limen (*DL*) for distance (d) in the *APS*.

The distance (d) between any two points specified in the *APS* may be given by the equation,

$$d = ((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)^{1/2} \tag{3.1}$$

where x_1, y_1, z_1 and x_2, y_2, z_2 are the coordinate values of the two points in *APS*. Data on difference limens (DLs) for formant frequencies of vowel sounds (Flanagan, 1955; Kakusho and Kato, 1968; Mermelstein, 1978; Nord and Sventelius, 1979) indicate that an average DL for d should be on the order of 0.02 log units. However, when extreme DL values for the first and second formant frequencies of vowel-like sounds from studies by Flanagan (1955) and Mermelstein (1978) are converted to *APS* coordinate values, they yield values for d which vary from as large as 0.0947 to as small as 0.0056 log units. Additionally, it should be noted that neither of the aforementioned studies varied $F3$ and only Mermelstein (1978) investigated simultaneous variation of formant frequencies as a sub-condition of one experiment. Thus, these findings may not reflect the difference limen found when more than one formant frequency is varied at a time. In his concluding remarks, Flanagan (1955) called for the experimental work to be detailed here, i.e., a mapping of DL areas on the $F1 - F2$ plane and the simultaneous variation of multiple formants, as a necessary extension of his work reported at that time, although, to the author's knowledge, no such work has been reported. Although parametrically varying the frequencies of up to three formants simultaneously is difficult to specify in a common metric, the *APS* provides an excellent format for such specifications since the complicated changes reduce to interpretable changes in location. Given consideration of these factors, we are motivated to determine the *DL* for d through our own investigation.

An additional relevant issue in speech perception research is whether or not discrimination ability is greater for vowel stimuli belonging to different phonetic categories as opposed to stimuli belonging to the same category. Fry, et al. (1962) found no significant differences between the discrimination of isolated vowels, whether from within a phoneme region or spanning a phoneme boundary. Stevens, et al.(1969), on the other hand, found distinct peaks in discrimination functions at vowel boundaries. Pisoni (1973) also found evidence

for differences in vowel discrimination of within-category versus between-category stimuli related to the comparison delay interval. Macmillan, Braida, Goldberg, and Khazatsky (1986) suggest, however, that many of the discrimination differences reported can be explained in terms of the stimulus range and psychophysical method employed and that these differences may not exist when proper stimulus parameters and methods are employed.

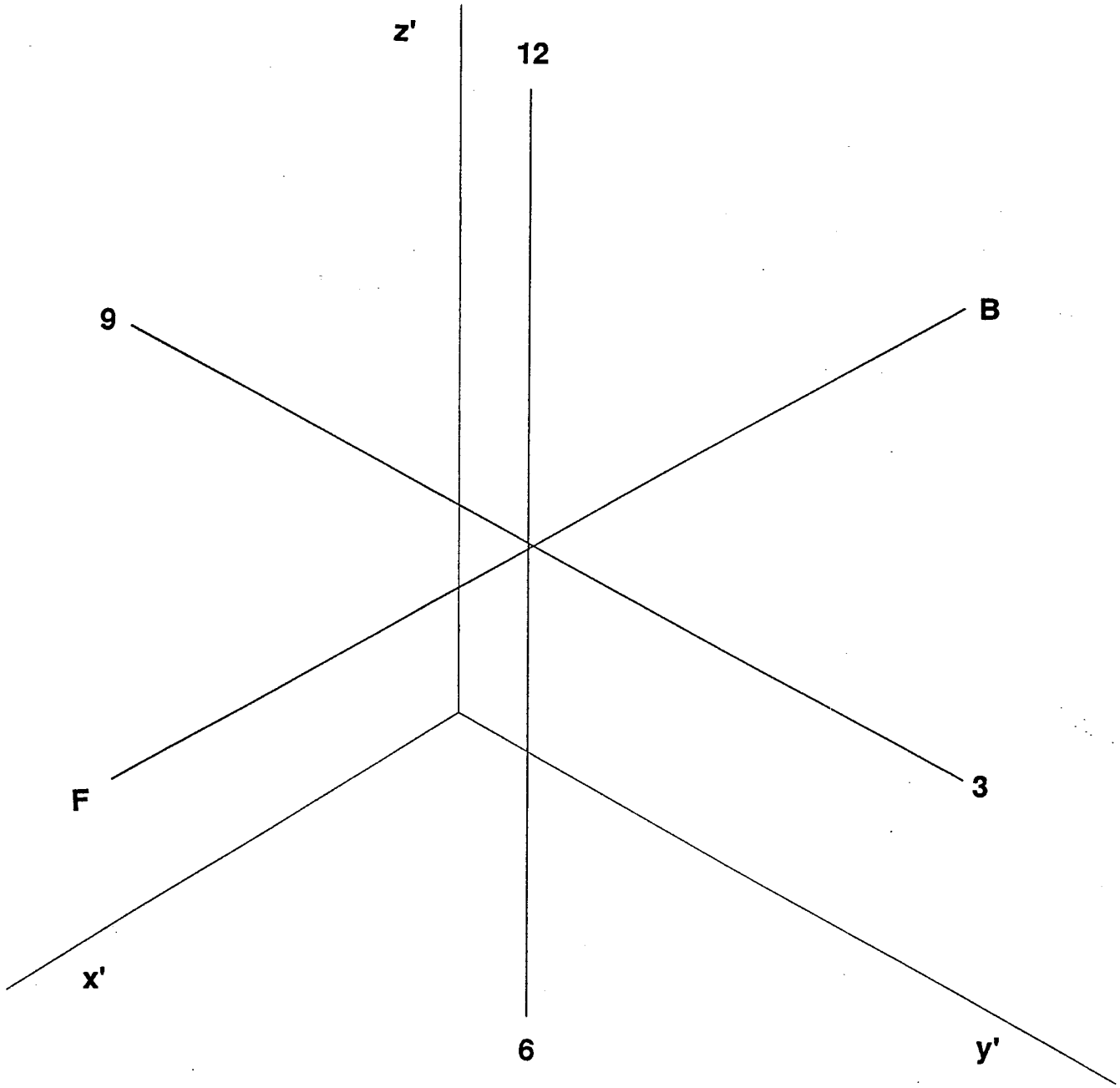
In Experiment I, a roving identification paradigm was used in an effort to restrict subjects to use only their own internal references for vowel sounds as a judgement basis for their responses. While this method does not yield as sensitive results as fixed discrimination paradigms (Macmillan, et al., 1986), it is considered adequate for the large range and relatively coarse resolution required in that experiment. However, a higher level of discrimination may be required for determining more exact boundary locations. Macmillan, et al. (1986), in generalizing the Durlach and Braida (1969) theory of intensity perception to speech perception, indicate that resolution performance is highest when 1) the stimulus range is small, 2) the inter-stimulus interval is small, and 3) a fixed discrimination paradigm is used, such as two-interval forced-choice (2IFC) or same-different (AX). The design of Experiment II reflects consideration of these factors.

3.2 Methods

3.2.1 Stimuli

Vowel tokens were synthesized using formant frequencies specified by points along continua in the *APS*. These continua represent straight lines in the *APS*, with groupings of three continua sharing a common center value, or *reference*. Each three-continua grouping is oriented so as to graphically form a three-dimensional "cross-hair", as shown in Figure 3-1, with each continuum of the crosshair parallel with one of the x', y', z' axes. Recall that these axes are transformations of the *APS* dimensions, x, y, and z (See Section 1.2.1). Each continuum is further subdivided at the reference point, such that the reference serves as a common point shared by each of what is now six continua. For convenience, the six continua are labeled in clock-like fashion for continua parallel to x' and y' (i.e., 12 o'clock= upward along y', 6 o'clock= downward along y', 3 o'clock= right along x', and 9 o'clock= left along

Figure 3-1: Orientation of the six continua in APS $x'y'z'$ coordinates associated with each reference point used in Experiment II.



x'), and F and B (i.e., front and back) for continua parallel to z' (See Figure 3-1).

Initially, a fixed length of 0.02 log units and step size of 0.00025 log units was evaluated for each continuum which yielded 80 tokens per continuum. However, as will be discussed later, the amount of formant change given a fixed step size varies with the axial orientation of the continuum. Thus, given that the synthesizer requires integer formant value specifications, the initial step size utilized sometimes yielded identical formant values for neighboring tokens. Additionally, early pilot studies revealed that the initial continuum length may be too short to provide adequate estimation of some difference limens. Based on these studies, the following specifications were utilized to alleviate these difficulties and aid in the efficiency of testing. Continua parallel to the x' axis were 0.02 log units long with 80 tokens spaced at 0.00025 log units, continua parallel to the y' axis were 0.04 log units long with 120 tokens spaced at 0.00033 log units, and continua parallel to the z' axis were 0.015 log units long with 60 tokens spaced at 0.00025 log units. Additionally, a computer program was implemented which evaluated the formant specifications for all tokens in a given continuum and eliminated any tokens which were duplicates. Although the continua were now of varying lengths and step sizes, these changes assured that the difference limens could be estimated within a reasonable range and that all tokens along any given continuum varied by at least 1 Hz in at least one of the first three formant specifications.

Token duration, $F0$ and amplitude contours, formant-bandwidth calculation, higher-formant values, and all other global synthesis parameters were identical to those previously specified for token synthesis in Experiment I (See Section 2.2.1).

Reference points were selected at 17 locations in the *APS*, 10 from the interiors of the target zones, [IY, IH, EH, AE, AA, AH, AO, UH, UW, ER] and 7 from estimated boundary areas between the target zones, [IY-IH, IH-EH, EH-AE, AE-AH, AH-AA, AH-UH, UH-UW]. This design yielded a total of 102 continua for evaluation.

Reference point selection criteria

Since this experiment began prior to the analysis of the results from Experiment I, locations for the 10 reference points from the interiors of target zones, hereafter referred to as *center* references, were selected based on the results of an earlier pilot study. This pilot study was

identical to Experiment I and utilized three subjects highly trained in phonetics. Tokens whose identifications had been twice unanimously agreed upon by the three subjects were first located for each of the 12 response categories used in Experiment I. If more than one token per category satisfied this criterion, the final token selection was made by the investigator. All center reference points were located in the $z' = 0.70$ plane except for [IY] ($z' = 0.75$) and [ER] ($z' = 0.55$). The locations of the center references ([ER] is not included) are graphically illustrated in x', y' space in Figure 3-2, along with the synthetic target zones for the $z' = 0.70$ plane from Experiment I. Additionally, the locations of the "EXEMPLARY" tokens (See Section 2.4.2) from Experiment I and the male averages from Peterson and Barney (1952) are shown for comparison. All points in the figure not originally located in the $z' = 0.70$ plane have been normalized in a manner previously described for Figure 2-18 in Section 2.4.4 such that their relative locations reflect only their values of $F1$ and $F2$.

Locations for the 7 reference points from boundary areas between zones, hereafter referred to as *ambiguous* references, were selected in a manner different from the center references. Gross areas of potentially ambiguous tokens in the $z' = 0.70$ plane for the seven boundaries were first selected based on identifications of synthetic vowels from a previous study (Miller and Hawks, 1989) described in Section 2.4.3. The tokens in these areas had received identifications that were mixed between the vowel categories they bordered, as well as generally low confidence ratings. Points for the tokens in these areas are shown in Figure 3-3 along with the target zones constructed for this z' plane from Experiment I. Further refinement of these areas was accomplished by examining the identifications of these same tokens from three new subjects, highly trained in phonetics. Once again, tokens receiving identifications mixed between the appropriate neighboring categories as well as low confidence ratings were noted. These tokens are shown in Figure 3-4. The final selection of reference points was made from these groups of tokens by the investigator and are shown in Figure 3-5.

Figure 3-2: Locations in APS $x'y'$ coordinates of center references (x's) utilized in Experiment II compared to locations of exemplary reference framework tokens (+'s, See Section 2.4.2) and male average vowels ($\star$'s) from Peterson and Barney, 1952.

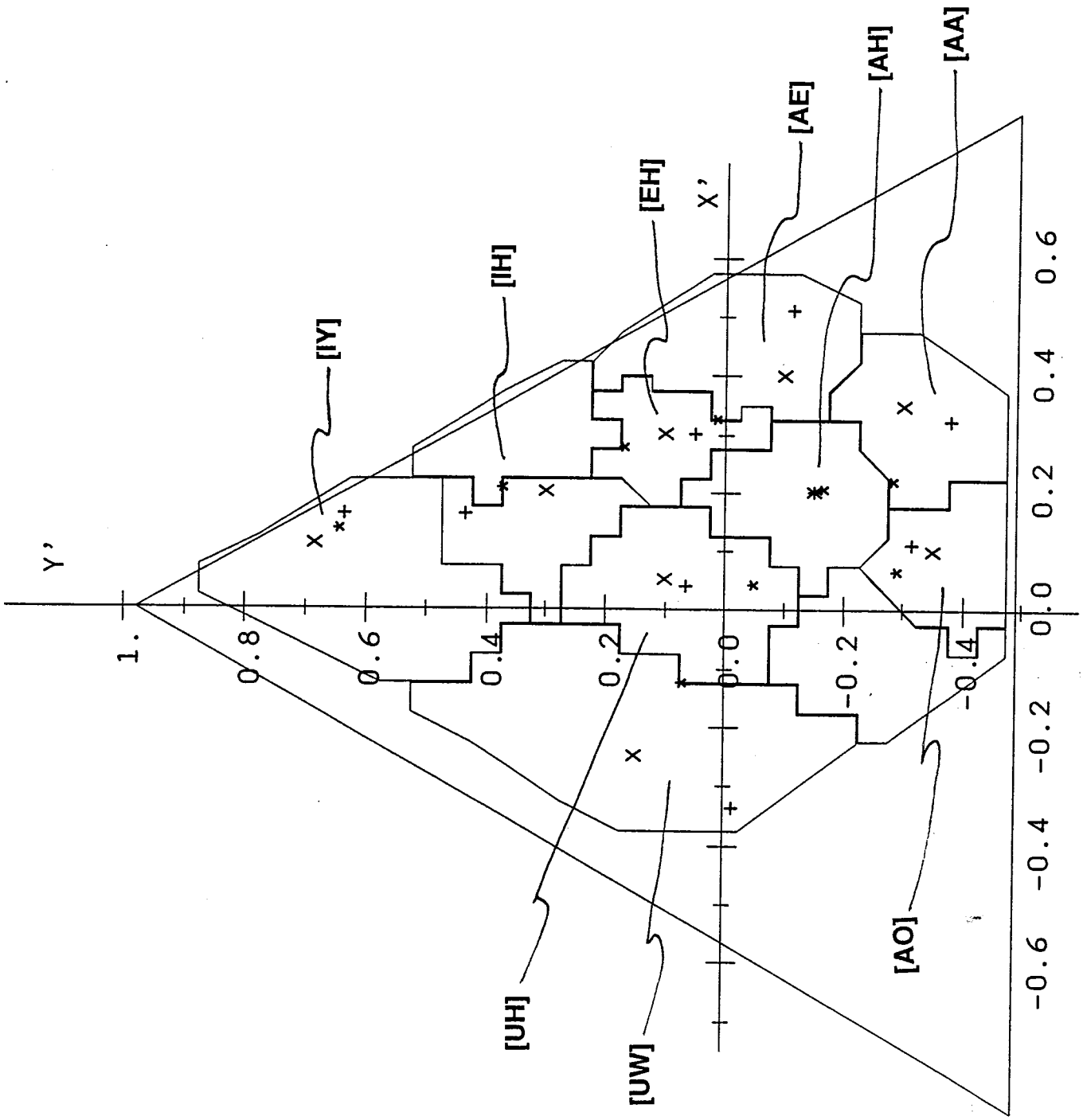


Figure 3-3: Locations in APS $x'y'$ coordinates of points for the first evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane.

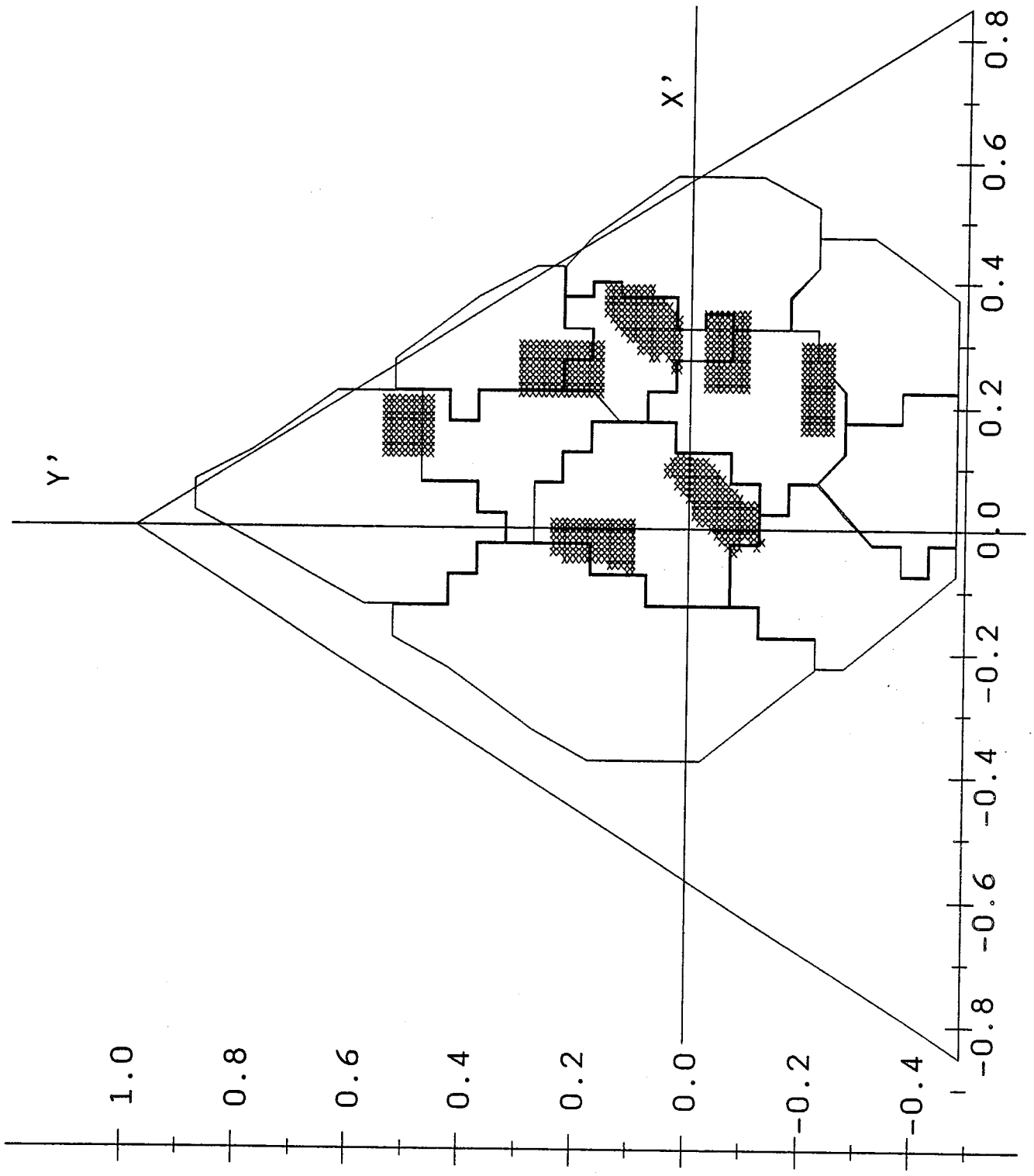


Figure 3-4: Locations in APS $x'y'$ coordinates of points for the second evaluation for ambiguous reference points along with *SSB* target zones for the $z' = 0.70$ plane.

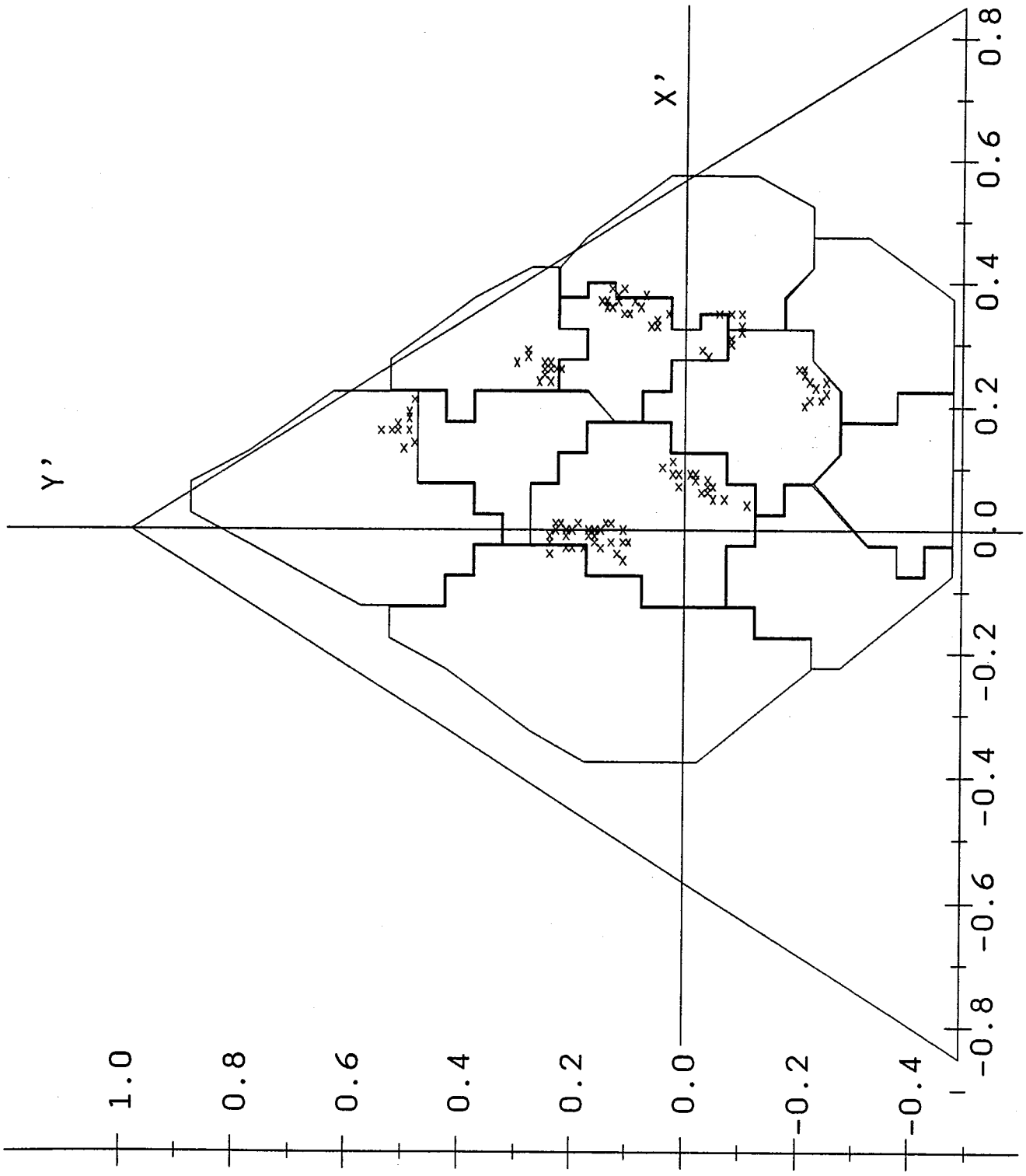
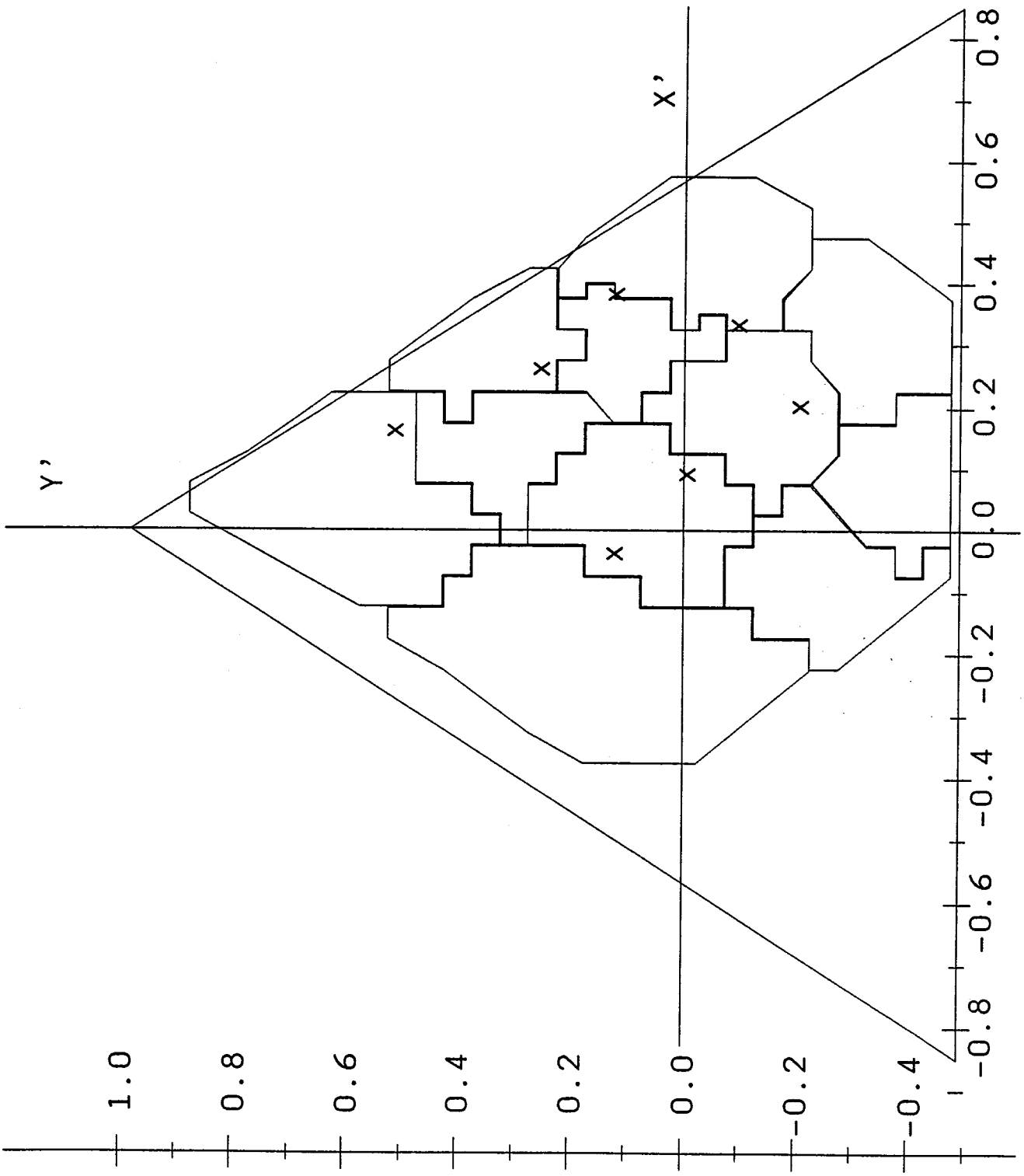


Figure 3-5: Locations in APS $x'y'$ coordinates of the ambiguous reference points used in Experiment II along with the *SSB* target zones for the $z' = 0.70$ plane.



3.2.2 Procedure

An adaptive up-down procedure similar to the PEST procedure (Taylor and Creelman, 1967) was utilized employing a cued, two-alternative forced-choice (2AFC) task. In this procedure, stimuli were presented in decreasingly smaller steps with the directional movement along the continuum determined by the subject's previous responses. The initial step size was 20 and approximately halved the first four times the direction along the continuum reversed (10, 5, 2, 1). Unlike the regular PEST procedure, no provision was made for increasing the step size. A 3-down, 1-up rule was also employed to the directional criteria. This rule allows movement toward the reference only after 3 correct responses at the current level and movement away from the reference after 1 incorrect response. This strategy should yield a probability of a correct response at the point of convergence of 0.794 (Levitt, 1970).

A fixed-standard vowel token was presented as a cue followed by two other vowel tokens, one of which was the same as the standard. The three vowel tokens comprising each trial were separated by inter-stimulus intervals of 250 msec. Subjects were asked to judge whether the different sound was in the first or second interval following the cue. The fixed-standard vowel tokens used represented the reference points of the cross-hairs. Thus, the six thresholds from each cross-hair were estimated by measurements made from the same reference point outward in all six directions. The response interval was subject paced, such that the next trial began 3 sec after a response was registered. Trial presentations ended, constituting a block of trials, after the subject had made 14 reversals along the continuum. Data from the first four reversals were discarded and the threshold was estimated as the average of the remaining 10 reversal points.

Subjects worked two hours per day and could complete about twelve blocks of threshold measurements within this time period. Blocks were randomized among the 17 reference points and the six directions for each subject. Each continuum was evaluated twice by each subject, yielding a total of 816 (17x6x4x2) threshold estimates.

Although same/different (AX) tasks have often been used for experiments of this type (Flanagan, 1955; Mermelstein, 1978; Nord and Sventelius, 1979), the cued, two-alternative forced-choice task is selected here in an effort to eliminate response criterion differences

between subjects. With a fixed-standard cue and the requirement of a judgement based on the comparison of two intervals, subjects are less likely to utilize an internal standard. A 250 ms delay interval has been found to yield the highest values of d' in discrimination experiments with vowels (Pisoni, 1973), and was therefore used here as the inter-stimulus interval.

3.2.3 Apparatus

The vowel tokens were synthesized at a 10 kHz sampling rate and stored on a DEC 3200 computer. The testing paradigm was implemented as an interactive program on the same computer which each subject ran independently. The program operated as follows. Prior to each trial, a warning flashed on the screen that a new trial would begin in 2 seconds. Following the trial presentation, subjects saw a question on the screen asking whether the different vowel sound was in the second or third interval. Subjects entered their response by pressing one of two labeled keys on the keyboard. A file was generated for each block of trials which contained general subject and block information, as well as a detailed description of the presentation parameters and subject's responses for that block of trials.

The stimuli were presented via a MicroTechnology Digisound-16 digital-to-analogue converter followed by a passive 5 kHz anti-aliasing filter. Subjects heard the stimuli binaurally over Sennheiser HD-430 headphones at a comfortable listening level (55-60 dBA-Slow SPL) in a sound-isolated room.

3.2.4 Subjects

Subjects, two male and two female, were recruited from the student body of Washington University and the nearby St. Louis area. Subjects' ages ranged from 17 to 25 years. All subjects were native speakers of American English with no known history of either speech or hearing impairment. All subjects were naive in terms of any formal phonetic training.

3.2.5 Training

Training consisted of a minimum of four runs each on 14 continua not used in the actual experiment. Subjects were rejected if, by the end of the four runs, the average DL for

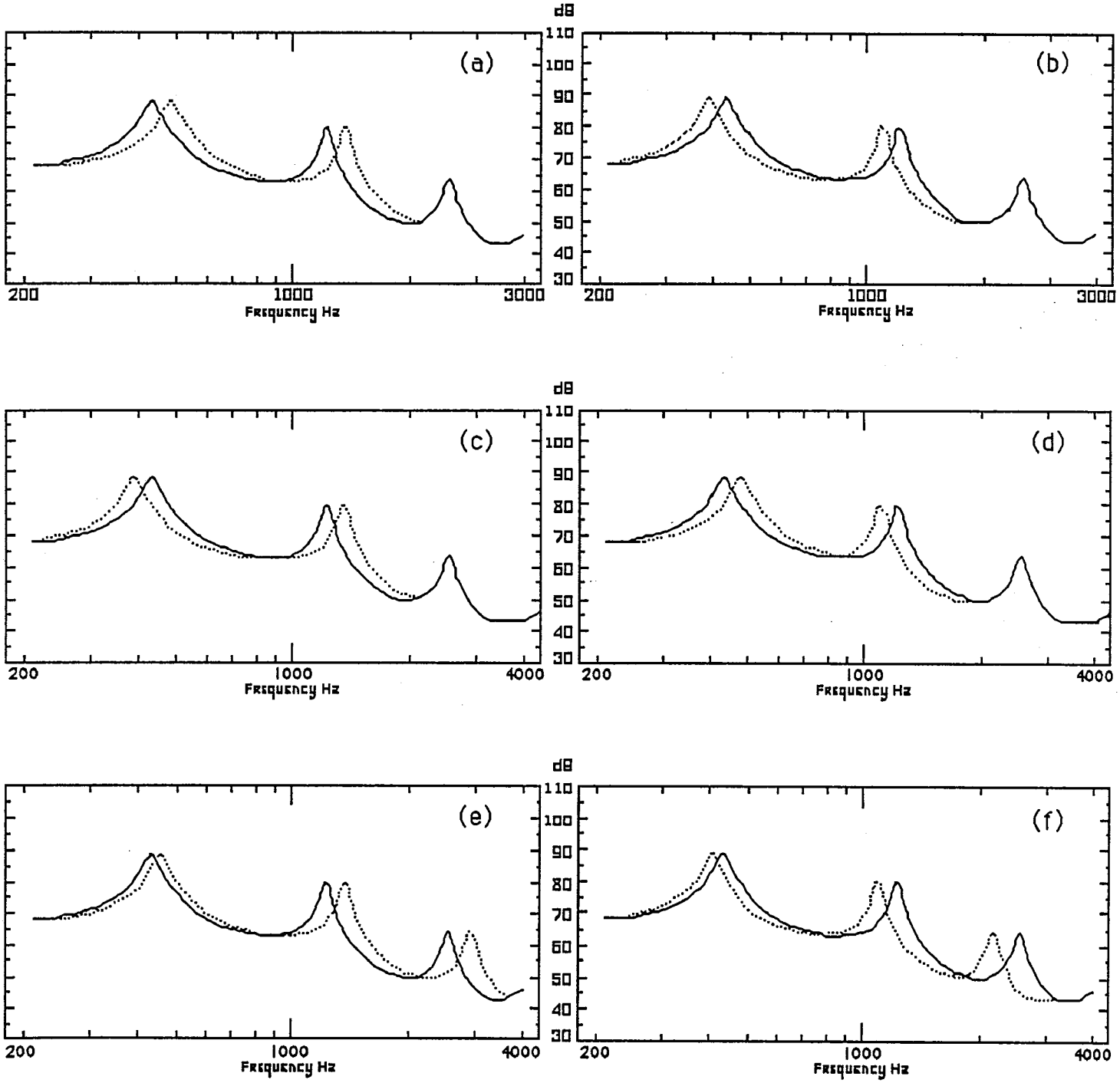
d over any one set of the 14 continua was not equal to or less than 0.01 log units. One subject was eliminated and replaced due to this criterion. The continua utilized in training included the six directions associated with each of the [OW] and [EY] center references and two additional continua, one from a central point in [AH] and one from the [AH-AO] boundary area. All other procedures and methods were identical to those used in the actual experiment.

3.2.6 Formant change and *APS* continua

To better describe the changes in spectral pattern under evaluation, this section will discuss the changes in formant frequencies derived from movement along the *APS* continua utilized in this experiment. First, consider points in *APS* at a fixed distance of 0.02 log units away from a cross-hair reference point along each of the six continua for that cross-hair. Given a fixed fundamental frequency, these six points and the reference point can be thought of as representing vowel sounds which differ systematically in two or all three of their specifications for $F1$, $F2$, and $F3$. These differences in formant specifications may be expressed as the differences between the logarithms of the formant frequencies represented by each of the six points along the continua and the logarithms of the formant specifications for the reference point. We will find that formant frequency shifts for $F1$ and $F2$ along some continua are often equal when expressed in this manner. However, a more traditional measure, $\Delta F/F$, where ΔF is the difference between the continuum point and the reference point formant frequency and F is the reference point formant frequency. This measure, expressed as a percentage, is also approximately equal for the relatively small changes we will be considering here and will therefore be used extensively for analyses of the results.

Continua 3 and 9 (See Figure 3-1) form a straight line through the reference point and lie parallel to the x' axis. Moving the fixed distance away from the reference point along continuum 3 yields positive changes in both $F1$ and $F2$, which are equal log differences between continuum and reference point formant frequencies, and approximately equal 3.3% changes in $F1$ and $F2$. Thus, for this continuum, $F1$ and $F2$ increase proportionately together. These differences are idealized graphically in Figure 3-6a as spectra, with the reference point represented with a solid line and the point along continuum 3 with a dashed

Figure 3-6: Idealized spectra representing the relative shifts in formant frequency for a fixed distance movement along each of the six directional continua (dashed lines) relative to the reference point (solid lines). (a) Continuum 3; (b) Continuum 9; (c) Continuum 12; (d) Continuum 6; (e) Continuum F; (f) Continuum B.



line[1]. Moving this same distance away from the reference point along continuum 9 yields an opposite effect (See figure 3-6b) whereby $F1$ and $F2$ both decrease proportionately by 3.3%. There is no difference between the $F3$ specifications of the reference point and points along these continua.

Continua 6 and 12 form a straight line through the reference point and lie parallel to the y' axis. Movement along these continua does not yield parallel changes in $F1$ and $F2$ as seen in continua 3 and 9, but rather, opposing changes, whereby $F1$ and $F2$ shift toward or away from one another an equal log differences between continua and reference point formant frequencies. For movement equal to the fixed distance away from the reference point along continuum 12, $F1$ will decrease by approximately 1.9% of the reference point $F1$ value and $F2$ will increase by the same percentage (Figure 3-6c). Once again the opposite effect is found for equivalent movement along continuum 6 (Figure 3-6d), whereby $F1$ will increase and $F2$ will decrease by approximately 1.9% of the reference point formant values. There is again no difference between the $F3$ specifications of the reference point and points along these continua.

For movement along the continua parallel to z', F and B, a more complicated pattern emerges. Movement equal to the fixed distance away from the reference point along continuum F results in an increase in all three formant values relative to the reference point values. $F1$ increases by approximately 2.8%, $F2$ by 5.5%, and $F3$ by 8.3% (Figure 3-6e). The reverse of these changes occurs for the fixed-distance movement along continuum B with all three formant frequencies decreasing by approximately these same percentages (Figure 3-6f).

This section has related how vowel sounds represented by points along the continua of a cross-hair vary in their formant frequency specifications from the specifications represented by the reference point for that cross-hair. From this discussion several important aspects of formant change with movement in *APS* emerge. First, the differences in formant specifications associated with points along a straight-line continuum which is parallel to one of the x',y', or z' axes in *APS* are systematic and may be calculated given knowledge of the

[1]The shifts indicated for formant frequencies between spectra in these figures has been exaggerated for visual clarity.

distance between the points and the axial orientation of the continuum. Second, the axial orientation of the continuum also determines the nature of formant change which results in distinctly different patterns of spectral change associated with each axis. For further elaboration on the relationships between movement in *APS* and formant changes, see Appendix A.

3.3 Results

3.3.1 General Analyses

Table 3.1 shows the mean DLs (rounded to two significant digits) as log unit distances from the reference point by reference point group, axis, and axial direction across subjects and replications. The overall average DL expressed as d across all continua was 0.0110 log units. The average DL for continua parallel to the x' axis was 0.0091 log units, the y' axis, 0.0190 log units, and the z' axis, 0.0035 log units. A repeated-measures ANOVA was computed on the experimental results employing the four subjects as replicates. A 2 x 3 x 2 x 2 factorial design was used for the analysis comparing the two reference point groups (center vs. ambiguous), the three axes (x', y', z'), the two axial directions (positive vs. negative relative to the reference point), and the two replications. Subjects' average DLs within each reference point group for each factor served as the analysis data.

The statistical probabilities of a significant difference for each comparison factor from this analysis are shown in the second column of Table 3.2. The most significant main effect found in the analysis was for differences between DLs when grouped by axis ($\rho = 0.0002$). Given the previous discussion of the differences in percent change of formant frequencies for a fixed-distance movement along these axes, if the DLs expressed as percent formant change for the three axes are relatively equal, these differences in DLs expressed as distance are not surprising. The only other significant main effect found was for replication ($\rho = 0.02$). Subjects demonstrated significantly better performance on replications of blocks, despite training and randomization of blocks. No significant difference was found between center and ambiguious reference point groups, although the average DL for the ambiguous reference point continua is smaller than the average DL for the center reference point continua.

Table 3.1: Mean DL in log unit distance for various conditions across subjects and replications.

Axis	Axial Direction	*Center* References	*Ambiguous* References	Row $\overline{x}$
x'	+	.0092	.0080	.0086
x'	−	.0110	.0083	.0095
x'	±	.0100	.0082	.0091
y'	+	.0180	.0200	.0190
y'	−	.0220	.0180	.0200
y'	±	.0200	.0190	.0190
z'	+	.0040	.0033	.0036
z'	−	.0040	.0028	.0034
z'	±	.0040	.0031	.0035
$\overline{x}$	+	.0100	.0100	.0100
$\overline{x}$	−	.0120	.0096	.0110
$\overline{x}$	±	.0110	.0100	.0110

Additionally, no significant difference was found between directions along the continua (e.g., continua 12, 9, and F vs. continua 6, 3, and B).

Table 3.2: Probabilities of significance for factors from overall and individual reference group analyses-of-variance of DLs expressed as distance.

Factor	Overall	*center* references	*ambiguous* references
Reference Group	ns	$\rho = .0017$	ns
Axis	$\rho = .0002$	$\rho = .0001$	$\rho = .0010$
Direction	ns	$\rho = .0170$	ns
Replication	$\rho = .0200$	$\rho = .0430$	ns

Significant interactions were found for axis-by-replication ($\rho = 0.012$), reference group-by-direction ($\rho = 0.026$), and reference group-by-axis-by-direction ($\rho = 0.033$). The axis-by-replication interaction seems to reflect the fact that while subjects' DLs improved across replications for the x' and y' axes, their performance remained the same across replications for the z' axis. The reference group-by-direction interaction stems from opposing differences in DLs between directions for the two reference point groups. The mean DL for increasing axis directions is smaller than for decreasing axis directions for center reference point continua, while the opposite is true for the ambiguous reference point continua. The three-way interaction between reference group, axis, and direction is more complex, but appears to predominantly reflect the reference group-by-direction interaction and the extremely strong main effect of axis.

3.3.2 Analyses by reference group

As was stated previously, no significant difference between the center and ambiguous reference points was found when they were considered as two distinct groups. However, that analysis did not evaluate possible differences within the groups. Separate ANOVAs were computed on the ten center reference points and the seven ambiguous reference points utilizing the same grouping factors of reference, axis, direction, and replication, and employing subjects as replicates. The statistical probabilities of a significant difference for each com-

parison factor from these analyses are shown in the third and fourth columns of Table 3.2. A significant main effect ($\rho = 0.0017$) for differences between the ten center reference points was found, however, a similar effect was not found for the seven ambiguous reference points. Significant differences for the axis factor were found for both reference point groups, however, differences for the direction and replication factors were significant only for the center references group.

Center reference points

The overall DL results, ranked in order from smallest to largest, for continua from each of the ten center reference points are shown in Table 3.3. Additionally, the DL results by

Table 3.3: Ranked DL results associated with each center reference point.

Vowel Category	Overall DL	x' DL		y' DL		z' DL	
IH	.0073	.0070	(1)	.0120	(1)	.0031	(3)
EH	.0087	.0083	(4)	.0150	(2)	.0025	(1)
AE	.0096	.0073	(2)	.0190	(4)	.0027	(2)
AH	.0097	.0090	(5)	.0170	(3)	.0032	(5)
AA	.0110	.0097	(6)	.0200	(6)	.0032	(4)
ER	.0110	.0080	(3)	.0200	(5)	.0063	(10)
UH	.0120	.0100	(7)	.0210	(7.5)	.0041	(7)
AO	.0130	.0130	(9)	.0210	(7.5)	.0055	(8)
UW	.0140	.0130	(8)	.0240	(9)	.0056	(9)
IY	.0150	.0140	(10)	.0290	(10)	.0037	(6)

axis are also shown for each reference in this table with rankings within that condition shown in parentheses. The rankings by axis agree reasonably well with the overall rankings with the exception of the reversed rankings for [IY] and [ER] for the z' axis. In general, discrimination appears best for the mid front vowels and worst for high vowels with low, central, and retroflex vowels intermediate.

Ambiguous reference points

Although no significant main effect was found for the ambiguous reference points, a rank ordering of the overall DLs and DLs by axis are shown in Table 3.4 for comparison. The

Table 3.4: Ranked DL results associated with each ambiguous reference point.

Vowel Category	Overall DL	x' DL		y' DL		z' DL	
EHAE	.0078	.0058	(1)	.0150	(2)	.0025	(3)
IYIH	.0085	.0094	(5)	.0130	(1)	.0027	(4)
IHEH	.0093	.0064	(3)	.0190	(4)	.0024	(2)
AHAA	.0100	.0095	(6)	.0170	(3)	.0037	(5.5)
AEAH	.0100	.0063	(2)	.0220	(7)	.0023	(1)
AHUH	.0110	.0079	(4)	.0220	(6)	.0037	(5.5)
UHUW	.0130	.0120	(7)	.0220	(5)	.0041	(7)

discrimination trend tends to follow that of the center reference points with better discrimination for references in the front vowel region and poorer discrimination for references in the back vowel region. The range for the overall DLs is somewhat smaller than for the center references. This may be due to the fact that the ambiguous points as a group tend to be more intermediate in terms of articulation and lack the extreme articulatory points which yielded the poorest DLs for the center references. Rank orderings between the axis results were considerably more varied than with the center references.

3.3.3 Analyses by Subject

To evaluate the uniformity of performance across subjects, an ANOVA was computed similar to the first general analysis previously described except subjects were now employed as the first grouping factor and replications served as replicates. A highly significant main effect for subjects was found ($\rho = 0.0006$). A large number of significant two- and three-way interactions were also found, suggesting that subjects were not only significantly different from one another, but that the ways in which they differed varied with the grouping factors.

Because of the varied individual differences between subjects, separate ANOVAs were

computed for each subject based on each subject's own performance across the various grouping factors once again employing replications as replicates. The significance levels for the main effects from these individual analyses are shown in Table 3.5 along with similar results from the overall analysis. Although the significance levels for individual subjects

Table 3.5: Analysis of variance results for effect of conditions overall and by subject.

Variable Grouping	Overall	Subject 1F	Subject 2F	Subject 1M	Subject 2M
Reference Group	ns	ns	ns	ns	***
Axis	**	***	***	*	**
Direction	ns	*	*	ns	ns

*** $p < 0.001$.
** $p < 0.01$.
* $p < 0.05$.
ns: $p > 0.05$.

should be viewed only as an indicator of which subjects strongly followed the direction of the overall results, obvious individual differences are apparent. While the results for three of the four subjects agree with the overall non-significant result for differences between center and ambiguous reference groups, the results for subject 2M (second male) indicate a high level of significance. Similarly, the non-significant results for direction from the two male subjects, 1M and 2M, agree with the overall results, but significant results are found for the female subjects, 1F and 2F.

3.3.4 Analyses by Percentage of Formant Change

As has been discussed previously, $\Delta F/F$ for a fixed distance in *APS* varies with axial orientation, therefore making comparisons of DLs as percentages in Hertz difficult, if evaluated as distances along different axes. To normalize across these differences and enable comparisons between axes, the DLs may be evaluated in terms of percent F change, i.e., ΔF expressed as a percentage of the reference formant frequency value, or $100(\Delta F/F)$, rather than distance.

A normalization of this kind is of particular interest for evaluating differences in DLs related to the x' and y' axes. While the frequency shifts for $F1$, $F2$, and $F3$ related to movement parallel to z' has been shown to be somewhat complicated, frequency shifts for

$F1$ and $F2$ related to movement parallel to x' and y' is similar. Changes parallel to either axis result in approximately equal shifts for both formants when expressed as percentages of F change with the difference between axes now being the pattern of spectral change. The differences between DLs related to direction along these two axes can be further normalized by utilizing the absolute values (i.e., disregarding the signs) for the percentages of F change.

To evaluate the difference between DLs associated with the two spectral patterns induced by movement parallel to the x' and y' axes, a repeated-measures ANOVA was computed employing the four subjects as replicates. The analysis was similar to that used for the first general analysis except that the absolute values of percent $F2$ change were used as the DL variable instead of log unit distance. Since percentages calculated from $F1$ change and $F2$ change are approximately equal for any given movement along each axis, these analyses could have been computed using percentages of $F1$ change which should have yielded the same results. A 2 x 2 x 2 x 2 factorial design was used for the analysis, comparing the two reference point groups (center vs. ambiguous), the two axes (x' vs. y'), the two axial directions (positive vs. negative axis), and the two replications.

The results of this analysis proved to be very similar to the results of the first general analysis discussed in section 3.3.1. There was no significant main effect once again for differences between center and ambiguous reference point groups ($\rho = .097$) or differences between positive and negative directions along continua. However, a significant main effect was found for differences between DLs associated with the x' and y' axes ($\rho = .033$) and for differences between replications ($\rho = .043$). A significant interaction between the axis and direction factors ($\rho = .034$) was also noted and follows the same pattern as was discussed in section 3.3.1. The significant effect of most interest from this analysis however, is the difference between x' and y' axes. The average percent $F2$ change for continua parallel to the x' axis was 1.47% and to the y' axis, 1.81%. This result suggests that there is a significant difference in how subjects discriminate the two different spectral patterns represented by movement parallel to these axes. Subjects seem to better discriminate patterns where $F1$ and $F2$ move in like directions (x') than patterns where $F1$ and $F2$ move in opposing directions (y').

Separate ANOVAs were once again computed to evaluate differences within the two

reference point groups using the percent $F2$ change DLs. The same factorial design was used for the analyses as was described above, except that the reference point factor now had 10 levels for the center reference point analysis and seven levels for the ambiguous reference point analysis. The results of these analyses are compared to the overall analysis in Table 3.6. The patterns of significant effects for the two analyses are quite disperate.

Table 3.6: Significances of factors from overall and individual reference group analyses-of-variance of DLs expressed as percent F2 change.

Factor	Overall	*center* references	*ambiguous* references
Reference Group	ns	$\rho = .0008$	ns
Axis	$\rho = .033$	ns	$\rho = .030$
Direction	ns	$\rho = .046$	ns
Replication	$\rho = .042$	ns	ns

As before, a significant main effect was found for differences between the ten center reference points, but no similar effect was found for differences between the seven ambiguous reference points. The axis factor was significant for the ambiguous reference group analysis, but not for the center reference group analysis. The opposite was true for the direction factor with a significant result found for the center reference point analysis, but not the ambiguous reference point analysis. Each analyses had one significant two-way interaction.

For the center references analysis the reference-by-direction interaction was significant ($\rho = .002$). This interaction may be due to considerable differences in discrimination by axial direction between references, although overall discrimination is significantly better for positive axial directions. For the ambiguous references analysis the reference-by-axis interaction was significant ($\rho = .011$). A similar reasoning can be applied here, that is, while overall discrimination is significantly better for the x' axis, this pattern is reversed for several of the individual reference points.

In summary, we find that, when the differences in DLs for continua parallel to the x' and y' axes are expressed as absolute changes in formant frequency relative to the reference, significant statistical factors remain basically the same as was found for the differences in

DLs expressed as distance. There is no significant difference between the center and ambiguous reference groups. Within these groups, however, the center references are significantly different from one another, but not the ambiguous references. Direction along axes was not found to be a significant factor overall, although significant differences exist for the center references with the better discrimination found for positive axial directions. Replication was a significant factor overall, but not at the individual reference group level. Perhaps of greatest interest, however, is that a significant difference was still found for discrimination between continua associated with the x' and y' axes. This difference is largely accounted for by significant differences for this factor within the ambiguous references, although the same pattern of difference, that is, continua associated with the x' axis are better discriminated than continua associated with the y' axis, is seen for the center references as well. The significant difference found between these axes also implies that there is a difference in how the two types of spectral pattern change associated with these axes, that is, parallel formant movement vs. opposing formant movement, are discriminated and that the better discrimination is for formant patterns exhibiting parallel movement.

3.3.5 Analyses of Movement in z'

As was stated earlier, movement along continua parallel to the z' axis represents unequal percent changes in $F1$, $F2$, and $F3$ making these continua less comparable to movement parallel to x' or y'. Therefore, three separate analyses have been computed on the log unit distance measures for DLs in z'. These analyses follow the same format as those for the x' and y' comparisons where the first analysis compares between center and ambiguous reference point groups, and the following analyses compare within the two groups.

The first analysis was a 2 x 2 x 2 factorial design ANOVA utilizing subjects as replicates and comparing center vs. ambiguous reference points, positive vs. negative directions along the axis, and the two replications. The within-group analyses followed the same design except that the first factor had ten levels for the center reference group and seven levels for the ambiguous reference group. The results of these analyses are shown in Table 3.7 and indicate that significant differences exist not only between center and ambiguous reference point groups, but also within each group.

Table 3.7: Analysis-of-variance results for effect of conditions overall and by reference group for z' continua.

Variable Grouping	Overall	*center* references	*ambiguous* references
Reference Group	$\rho = .024$	$\rho = .0008$	$\rho = .011$
Direction	ns	ns	$\rho = .025$
Replication	ns	ns	ns

For the between-group analysis, the ambiguous references as a group were discriminated significantly better than the center references. No significant differences were found for direction or replication. Significant interactions were found for group-by-direction ($\rho = .048$) and group-by-replication ($\rho = .024$). The group-by-direction interaction may be the result of significantly different discrimination for direction found among the ambiguous references interacting with virtually no difference in discrimination for direction among the center references group. The group-by-replication interaction may reflect the product of opposing trends for replications between the two groups.

The results of the within-group analyses are similar to the between-group analysis except that the direction factor is significant for the ambiguous reference continua with negative-going continua the better discriminated. No significant interactions were found for either of these analyses. The rank ordering of DLs by reference along the z' axis were shown previously in Tables 3.3 and 3.4.

3.3.6 Overall Discrimination by Reference

Examination and discussion (See Section 3.3.2) of Tables 3.3 and 3.4 indicated that the DLs associated with individual references can vary considerably for any of the three axes. Individual formant locations and patterns of $F1$, $F2$, and $F3$ may be considered as a plausible explanation for differences in the DLs in general between the various reference points. The average percent $F2$ change ($100(\Delta F2/F2)$ for all x' continua by reference point are shown in Figure 3-7 along with the frequencies of $F1$, $F2$, and $F3$ of the reference points. A similar illustration for all y' continua is shown in Figure 3-8. Values for like

Figure 3-7: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for x' continua.

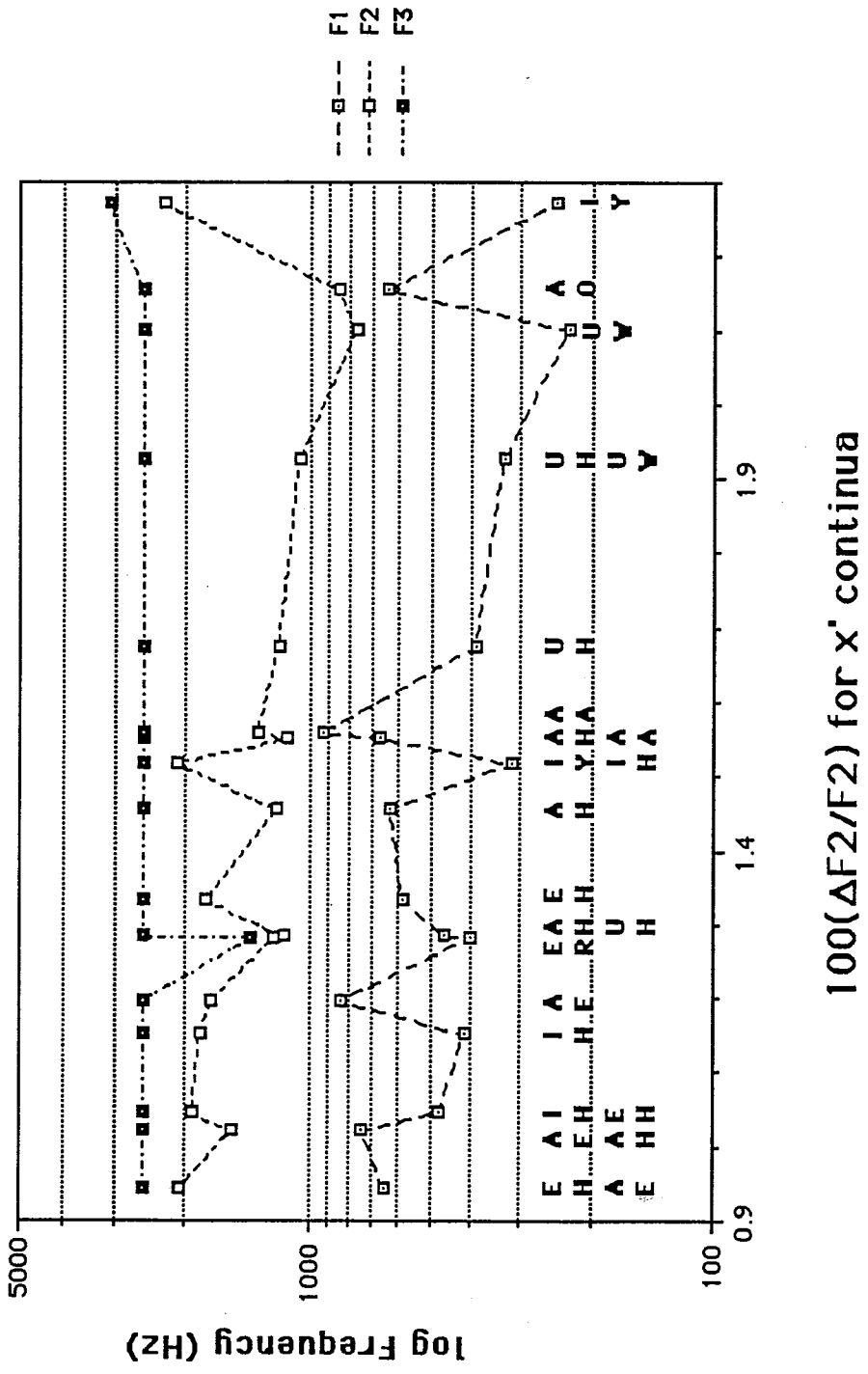
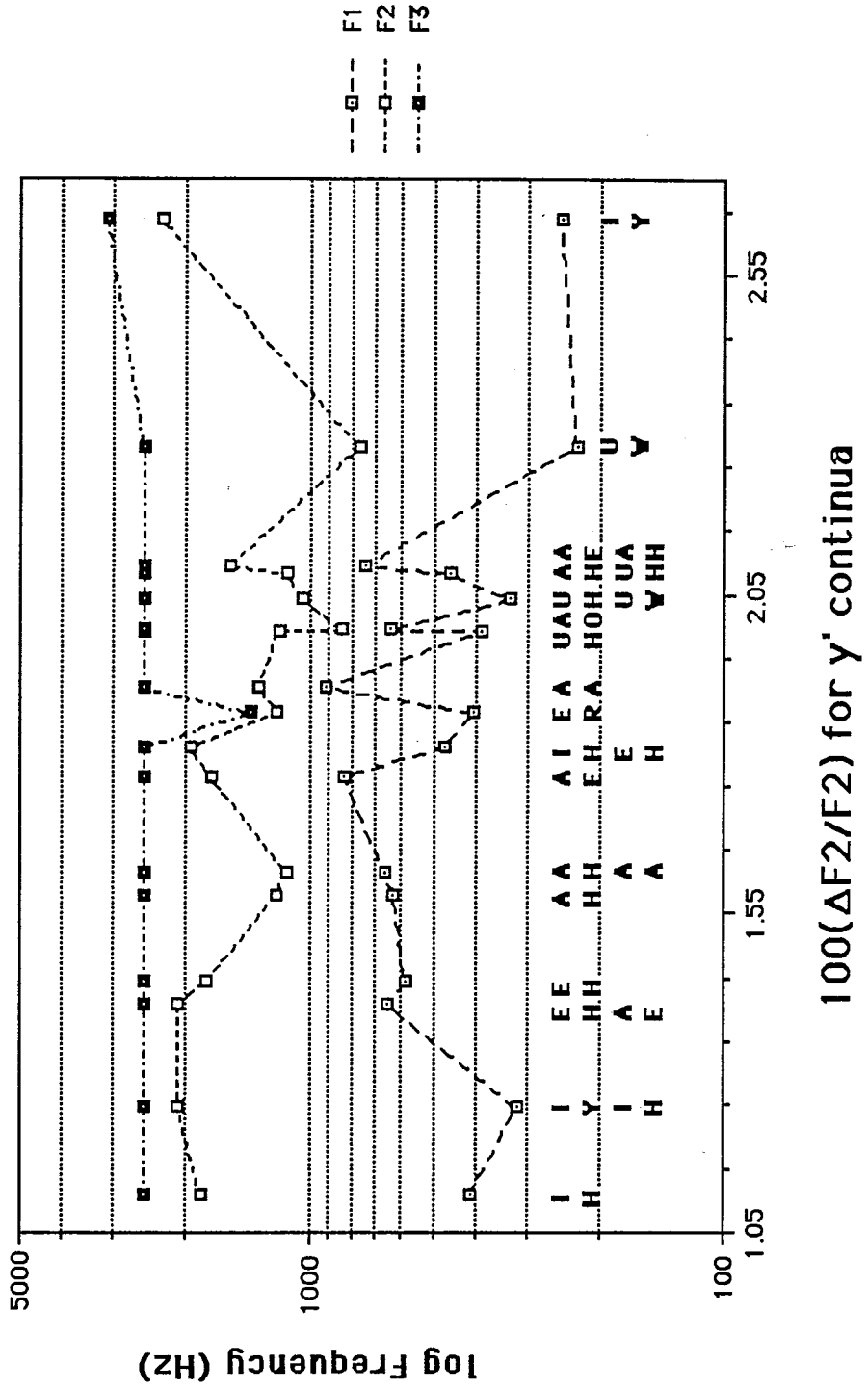


Figure 3-8: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for y' continua.



formant positions are connected by broken lines for ease of illustration. For the x' continua in Figure 3-7 we find a general trend for smaller DLs to be associated with higher values of $F2$, and with DLs increasing with decreasing values of $F2$. A pattern associated with $F1$ is less discernible but, in general, given a value of $F2$, lower values of $F1$ are associated with smaller DLs and higher values of $F1$ with larger DLs. Exceptions to both of these patterns are DLs associated with the /IYIH/ and /IY/ reference points. For these cases, we must assume that the extreme distances between $F1$ and $F2$ reduce discriminability.

Examination of Figure 3-8 for the y' continua reveals a somewhat similar pattern. The formant patterns for reference points associated with the best and worst DLs seem to generally follow the trends noted for the x' continua, but the majority of reference points lie intermediate to these and exhibit considerable variation in formant pattern.

The ordering of reference point formant patterns for z' continua, as seen in Figure 3-9, exhibits the same trend for decreasing $F2$ with increasing DL as seen for the x' and y' continua. However, the trend in the $F1$ pattern seems to be reversed from the $F1$ pattern seen in the previous figures, with $F1$ now decreasing with increasing DL, paralleling the $F2$ trend. Notable exceptions are found for /IY/, which is better discriminated along this axis, and /ER/, which, presumably due to the close proximity of $F2$ and $F3$, is the most poorly discriminated reference.

3.3.7 Single vs. Multiple Formant Movement

In addition to the continua previously specified which reflected multiple simultaneous formant changes, eight other continua reflecting single formant changes were also evaluated twice by three of the four subjects. These continua (see figure 3-10) represent straight lines in *APS* which emanate at 60 (labeled continuum 2), 120 (continuum 4), 240 (continuum 8), and 300 (continuum 10) degrees relative to the x' axis from the [EH-AE] and [AH-UH] ambiguous reference points. Points along the 60° and 240° continua represent vowel tokens where only $F1$ varies relative to the formant values of the reference point. Points along the 120° and 300° continua represent vowel tokens where only $F2$ varies relative to the formant values of the reference point. Points located 0.02 log units away from the reference point along these continua represent changes of approximately $\pm$ 3.3% in $F1$ or $F2$ relative to

Figure 3-9: Formant frequencies ($F1$, $F2$, and $F3$) in log Hz for all reference points (vertically labelled below each formant set) ordered by mean DL expressed in percent $F2$ change for z' continua.

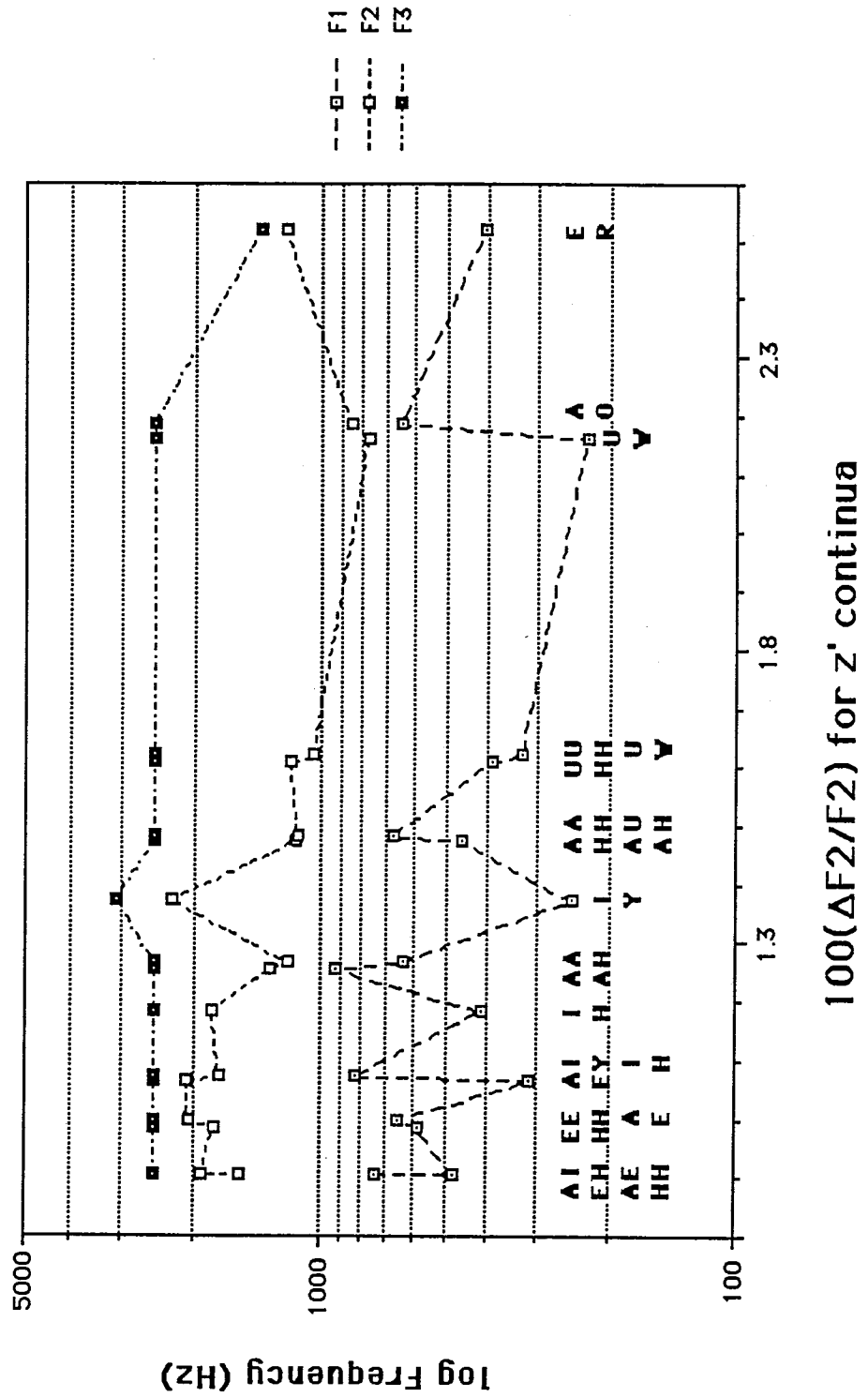
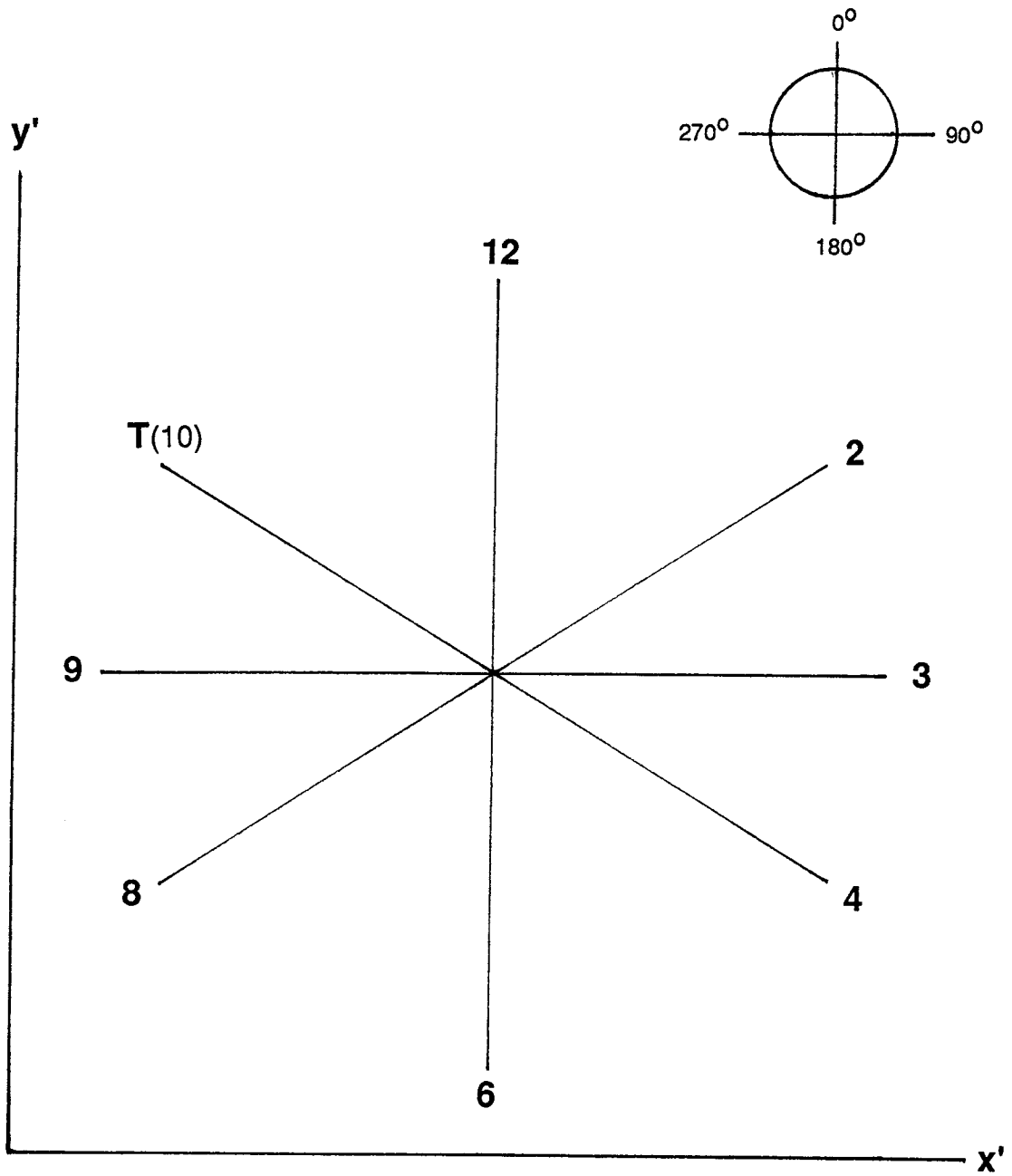


Figure 3-10: Locations in APS $x'y'$ coordinates of single-formant-change continua relative to multiple-formant-change continua.



the reference point formant frequencies, similar to the percentage of formant change found for points at this distance along continua parallel to the x' axis. No evaluation was made for changes in $F3$ alone.

A single-factor ANOVA indicated that differences in percentage of formant change between $F1$ and $F2$ were not significant. Therefore, the DL estimates along continua exhibiting single formant changes were directly compared to the DL estimates from continua parallel to the x' and y' axes exhibiting multiple formant changes emanating from the same reference points. Several questions of interest may be addressed by these comparisons. First, are DLs for single-formant-change continua significantly different from DLs for multiple-formant-change continua? Secondly, are DLs for single-formant-change continua more like the DLs for continua parallel to x' or y'?

The comparisons were once again based on DLs expressed as absolute percentages of $F1$ and $F2$ change as was done previously for comparison of x' and y' continua. The first comparison was a 2 x 2 x 2 x 2 x 2 factorial ANOVA utilizing the three subjects as replicates. The factorial design compared by group (i.e., single- vs. multiple-formant-change), by reference (AHUH vs. EHAE), by continua (continuua 2, 4, 8, and 10 vs. continua 3, 6, 9, and 12), by direction (continua 2, 4, 3, and 12 vs. continua 8, 10, 6, and 9), and by replication. The results of this analysis indicated only one significant main effect which was for the by-group factor ($\rho = .023$). This effect indicated that the DLs for the multiple-formant-change continua were significantly smaller ($\overline{x} = 1.42\%$) than the DLs for the single-formant-change continua ($\overline{x} = 1.89\%$). No significant interactions were found. This analysis was repeated with the groupings for the by-continua factor reversed such that continua 4 and 10 were grouped with the y' continua and continua 2 and 8 were grouped with the x' continua. This analysis yielded virtually the same results. Thus the results of these analyses suggest that the answer to the first question is yes, DLs for single-formant-change continua are significantly different from DLs for multiple-formant-change continua.

Recall that movements along the x' and y' axes represent two distinctly different changes in spectral pattern and that significant differences exist between DLs for these two axes. Although we find that DLs for single-formant-change continua are significantly different from the DLs for the x' and y' axes when grouped together, it is of interest to know whether

discrimination of changes in spectral patterns associated with varying a single formant is different from discrimination of changes in the two spectral patterns associated with the x' or y' axis (i.e., parallel or opposing shifts in $F1$ and $F2$) or more similar to one of them. To address this question, analyses of DL values expressed as absolute percentages of $F1$ or $F2$ change for the single-formant-change continua were compared first with corresponding values from continua parallel to the x' axis and then similarly with the y' axis. These comparisons were by way of ANOVAs using the same factorial design as for the first analysis and the same groupings for the first two factors (i.e., single- vs. multiple-formant-change and AHUH vs. EHAE). However the third grouping factor now becomes $F1$ vs. $F2$ which groups DLs for continua 4 and 10 with x' or y' $F1$ DLs and continua 2 and 8 with x' or y' $F2$ DLs. Furthermore, grouping for the direction factor was now by positive or negative formant change and not by direction in the space.

These analyses indicated that the DLs for single-formant-change continua are significantly different from the x' DLs ($\rho = .017$) but not significantly different from the y' DLs ($\rho = .477$). The DLs for the x' axis are significantly smaller (($\overline{x} = 1.03\%$) than the DLs for either the y' axis ($\overline{x} = 1.81\%$) or the single-formant-change continua ($\overline{x} = 1.89\%$). This result suggests that the changes in spectral pattern being discriminated along single-formant-change continua are more like the spectral pattern changes associated with y' continua than with x' continua.

3.4 Summarization and Discussion of Experiment II

In summary, we find that, based on the results of this experiment, the average DLs for the x', y', and z' axes are approximately .01, .02, and .004 log units respectively. As applied to the pilot mapping of the $z' = 0.70$ plane in .01 log unit steps presented at the end of Chapter 2, these results suggest that such a grid size should reflect sufficient resolution to estimate vowel category boundaries accurately along x' and y' and represents a good compromise in resolution if a single grid size is dictated. However, we can expect redundant boundary information concerning tokens along y', and, had additional planes been mapped at .01 log unit resolution, insufficient boundary information pertaining to tokens along z'.

If a variable grid size were to be used for mapping, the number of tokens required could

be cut in half if the grid size along y' were increased to .02 log units. The gain in efficiency brought about by this token reduction would be lost however, if planes of tokens along z' must be mapped at .004 log units. A reasonable alternative may be to consider mapping only the z' area associated with the perception of retroflection since it has been shown in Experiment I and elsewhere that changes in $F3$ affect the perception of non-retroflex vowels very little, at least for American English.

Results from this experiment did not reveal significant differences in discrimination between reference points located within phonetic vowel categories and reference points located at phonetic boundaries. Investigation of the possible differences in perceptual discrimination of synthetic vowel sounds from within and between phonetic categories has been active since 1962, when Fry and colleagues concluded that there were no differences in discrimination for the case of isolated synthetic vowel sounds. Since that time a number of investigators have attempted to detail more specifically how and why vowel (and consonant) discrimination may vary relative to phonetic classification. Among the variables which have been investigated are isolated vowels vs. consonant-bound vowels (Stevens, 1968; Sachs, 1969; Mermelstein, 1978), duration (i.e., short vs. long) (Sachs, 1969; Pisoni, 1973; Mermelstein, 1978), and psychophysical methods (Sachs, 1969; Pisoni, 1973; Macmillan, et al., 1986; Macmillan, Goldberg, and Braida, 1988). Taken as a whole, this past research would suggest that differences in discrimination related to phonetic boundary proximity are most likely to occur with synthetic vowels of short duration, vowels in a consonant context, and paradigms utilizing a roving ABX or same/different (AX) task. The present experiment found no such differences and, given that the experiment utilized synthetic vowels of reasonably long duration in an isolated context and were measured with an adaptive, cued-2IFC task, we should not expect to find significant differences in discrimination related to phonetic boundary proximity.

A more perplexing issue pertains to the significantly smaller DLs found in this experiment as opposed to similar experiments in the past. In general, the DL for changes in $F1$ or $F2$ when ΔF is expressed as a percentage of the formant frequency has consistently been found to be on the order of 3% to 5% (Flanagan, 1955; Kakusho and Kato, 1968; Mermelstein, 1978; Nord and Sventelius, 1979). For multiple simultaneous formant changes, this

experiment found average DLs to be approximately half that, on the order of 1.5% to 1.8% of the formant frequency. This result in and of itself would be acceptable, if varying multiple formant frequencies simultaneously produces additive information which in turn induces an increase in discrimination. Mermelstein (1978) proposed a weighted additive model based on $\Delta F1$ and $\Delta F1$ for single formant variation to predict the increased discrimination he reported for simultaneous variation of $F1$ and $F2$. Additionally, Carlson, Granström, and Klatt (1979) found slightly increased perceptual distances for simultaneous variation of $F1$ and $F2$ relative to perceptual distances for single formant variation. The difficulty here lies in the fact that DLs found in this experiment for single formant variation, while shown to be significantly different from the multiple-formant-variation DLs, are still on the order of 2%, definitely less than results from past research would suggest.

A number of factors may be considered as potential explanations for the differences in DLs from the present experiment and similar work in the past. All of these factors are related to the psychophysical methods and experimental protocols employed. The first consideration is subject training. Neither Flanagan (1955) nor Nord and Sventelius (1979) mention any training for their subjects and Mermelstein (1978) reports that subjects listened to about ten pairs of stimuli as familiarization prior to testing. In the present experiment, each subject was trained with a minimum of four runs on 14 different continua, about 3400 trials, prior to testing. Despite this degree of training, subjects still demonstrated significant improvement in their discrimination performance with replication. Although Flanagan (1955) utilized five repetitions and four subjects and Nord and Sventelius (1979) an average of 39 repetitions and 27 subjects, only Mermelstein (1978) makes mention of any significant differences between subjects. It is most probably safe to assume that subjects were better trained in the present experiment than in the studies mentioned above and this difference could yield more sensitive discrimination results.

Next, we consider the differences in the experimental methods used to estimate DLs as a possible factor for differences in discrimination. All three experiments cited from the past utilize a two-interval, same/different (2IAX) task from which the percentage correctly identified as "different" is plotted and the 50% correct point estimated. Although not explicitly stated in all three vowel DL studies cited, this task is often designed in a roving

discrimination pattern, i.e., any two stimuli among the references and their variations may be paired randomly, and results are corrected for guessing based on assumptions underlying "low-threshold" theory interpretations of discrimination experiments proposed by Luce (1963). The present experiment, on the other hand, utilizes a two-interval, forced-choice (2IFC) task with a cued, fixed standard and is interpreted in terms of signal detection models. Macmillan, Kaplan, and Creelman (1977) and Macmillan, Goldberg, and Braida (1988) both investigated the implications of these two methods on discrimination in speech research. Both concur that with the 2IAX task, response bias tends to be large, listening strategies non-optimal, and performance substantially lower than performance for yes/no (YN) or 2IFC tasks. The 2IFC task, on the other hand, is relatively free of response bias and a more optimal method of estimating true sensitivity. Additionally, Macmillan, et al. (1988) demonstrated that fixed discrimination, like that used in the present experiment, is more sensitive than roving discrimination, most likely to due to decreased memory requirements and/or lower stimulus uncertainty. The inter-stimulus interval (ISI) also plays a larger role in roving discrimination designs (Pisoni, 1973). Pisoni found that a .5 sec ISI (the past DL experiments cited used .4-.5 sec ISIs) yielded somewhat lower discrimination sensitivity than the .25 sec ISI (used in the present experiment) for synthetic vowels of longer (300 msec) duration.

The present experiment employs a cue or standard (the reference token) preceding each observation interval. Khazatsky (1985), in modeling sensitivity changes, demonstrated that standards can improve performance, particularly in the region of the standard, for identification tasks. The use of trial-by-trial cues in discrimination tasks is somewhat less clear, although cues have been found to yield improved performance (Greenberg, 1962). The general thinking is that cues or standards help reduce subjects' uncertainty and reduce the influences of internal references. In overview, Robinson and Watson (1972) state that "... performance can be improved by providing the listener with as much information as possible about the to-be-detected signal." (p. 111).

The last methodological issue to be raised is the use of adaptive and non-adaptive testing paradigms. Kollmeier, Gilkey, and Sieben (1988) evaluated several adaptive staircase rules with several psychophysical procedures (2AFC and 3AFC) and found that while modelings

of the various tasks suggest that results from adaptive and non-adaptive techniques should be equivalent, human data indicated that the adaptive techniques tended to yield lower thresholds. Although empirical evidence is scarce, it is also generally held that subjects are able to learn a task more rapidly when adaptive methods are employed.

Taken together, the differences in methodology discussed here may well explain the differences in the results between the present experiment and similar past work. Subjects in the present experiment were probably better trained and numerous considerations were given to the experimental protocol to minimize uncertainty and maximize discrimination sensitivity.

A question of interest arises in consideration of the differences between formant DLs for single- and multiple-formant variation. Are the DLs related, that is, can multiple-formant DLs be predicted from single-formant DLs? The model proposed by Mermelstein (1978) considers changes in $F1$ and $F2$ to independently contribute information to discriminability of stimuli where both formants are simultaneously varied. In the model, the composite ΔF for these stimuli is calculated by

$$(\omega_1(\Delta F1)^2 + \omega_2(\Delta F2)^2)^{1/2} \tag{3.2}$$

where ω_1 and ω_2 are unspecified weighting factors. Mermelstein implies that the weighting factors are related to the relation of $\Delta F1$ and $\Delta F2$, however, our own attempts to replicate his predictions from his DL results have failed. If the weighting factors should be related to the percent change of each formant, then equal weightings of 1 may be used to predict the present data. The resulting predictions for the average multiple-formant-change DLs from the average DLs for single-formant change are quite accurate, within 2 Hz, of the actual average results. Thus it would appear that $F1$ and $F2$ may equally contribute information for discrimination. If the model were to be expanded to include $F3$, with $F3$ also contributing information on an equal basis, we could use the equation

$$(\omega_1(\Delta F1)^3 + \omega_2(\Delta F2)^3 + \omega_3(\Delta F3)^3)^{1/3} \tag{3.3}$$

to estimate changes for continua associated with z' in the present experiment. Since a DL for single-formant variation in $F3$ has not been estimated presently or in the past literature, we shall, for the moment, assume the prediction equation is accurate and solve for the missing

single-formant-variation DL value for $F3$. If weightings are assigned to reflect the relative percent formant changes resulting from movement parallel to z' ($\omega_1 = .337$, $\omega_2 = .663$, and $\omega_3 = 1$), the resulting DL for single-formant variation in $F3$ should be about 1.87% of the reference $F3$. While, intuitively this value seems small, it is in line with the single-formant-variation DLs found for $F1$ and $F2$, and suggests that such a model may have merit.

Flanagan (1955) found that DLs for single formant manipulation were affected by the relative proximity of two neighboring formants. He reported asymmetries in the positive and negative frequency DLs, particularly for $F2$, with some formant combinations. In general, DLs decrease for the direction toward a neighboring formant in close proximity and increase for the direction away from the adjacent formant. Flanagan suggested that these asymmetries were due to larger increases in formant amplitude with closely neighboring formants than with formants exhibiting larger spacing in frequency. Mermelstein (1978) found a similar pattern in his results. Nord and Sventelius (1979) likewise found asymmetries in their replication of Flanagan's work, but in the opposite direction, and were forced to reject Flanagan's conclusions regarding formant intensity relations. They noted that for small shifts in $F2$ (i.e., < 50 Hz), there was no drastic change in level for the $F2 - F3$ complex and suggested that perhaps the increase in intensity build-up will not appear unless the proper auditory analysis is made. Additionally, they found good correlation between their discrimination curves and spectral distance measures (described by Plomp, 1970) based on 1/3 octave band filter analysis.

For the present experiment, significant differences between axial directions were found only among center references for the x' and y' axes and among ambiguous references for the z' axis. In general, positive axis directions yielded better discrimination for the x' and y' axes and the negative direction yielded better discrimination for the z' axis. These results are difficult to compare directly to past results noted above in that positive and negative frequency DLs are confounded for continua parallel to y', where $F1$ and $F2$ move in opposing directions. Despite this difficulty, closer examination of the results indicates no specific pattern related to direction of formant frequency change and formant proximity. The lack of differences related to direction of formant frequency change in the present

experiment may, once again, be attributable to the differences in methodology previously discussed.

Section 1.2.6 outlined general trends seen in formant relations potentially related to differences in discrimination for continua associated with the various reference points. A predominant trend was that discrimination ability seemed to decrease with decreases in $F2$. While patterns in $F1$ related to discrimination ability were more variable with axis and more difficult to discern, it is clear that $F1$ also plays some role in the general discriminability of the various reference points.

Another possible avenue of explanation for discrimination differences between references is through analysis of their auditory spatial-frequency patterns. Spatial frequency, as it relates to vowels, considers the relation between two adjacent peaks in the spectrum as one modulation of a frequency expressed in cycles/octave. Likewise, the spatial frequency measured between $F0$ (or a reference) and $F1$ can be thought of as a relative location measure for the spectrum. Although an extensive analysis regarding spatial frequency will not be undertaken here, a preliminary investigation suggests that these measures may be great interest.

To further explore the possible accountability of discrimination differences between reference points related to the three spectral patterns mentioned previously, linear multiple regressions were calculated. Three sets of independent variables were used. The first set contained the frequencies for $F1$, $F2$, and $F3$ of the seventeen reference points. The second set contained the the x, y, and z values for the reference points. Recall that these values are log ratios of formants, where $x = log(F3/F2)$, $y = log(F1/SR)$, and $z = log(F2/F1)$. The third set contained the reciprocals of the x, y, and z values which are related to the spatial frequencies of the formant relations. The dependent variables for these analyses were the DL results expressed as percentage change in $F2$ (i.e., $100(\Delta F2/F2)$ for continua associated with the x', y', and z' axes. The results of these analyses are shown in Table 3.8.

For the DLs associated with x', all three variable sets are able to account for a significant proportion of the variance. However, the reciprocals of the log ratios of formants, related to spatial frequency, account for the greatest proportion and are highly significant ($\rho = .0001$).

Table 3.8: R^2 values from multiple regression analyses of DL results and specified variable sets (See text).

Variable Set	DL x'	DL y'	DL z'
Formants	.565	.248	.744
Log ratios	.605	.256	.832
Reciprocals	.788	.390	.467

Within the independent variable coefficients, the reciprocal of $log(F1/R)$ is the most highly significant ($\rho = .0001$), but the reciprocal of $log(F2/F1)$ is also significant ($\rho = .003$). This finding suggests that differences between references for these continua may be potentially accounted for primarily by the the relative location of $F1$ and secondarily by the relation of $F2$ to $F1$.

For the DLs associated with y', none of the three independent variable sets are able to significantly account for the variance. However, of the three sets, the reciprocals come the closest ($\rho = .084$). Once again, the ordering of independent variable coefficients indicates the reciprocal of $log(F1/R)$ accounts for largest amount of variance alone. The results suggest that none of formant-related variables considered can adequately account for DL differences in references for the y' continua.

For the DLs associated with z', all three independent variable sets are able to significantly account for the variance at the $p < .05$ level, however, the simple formant ratios provide the most significant accounting ($\rho = .0001$). The ordering of accountability within the independent variables remains as seen for the other axes, i.e., $log(F1/R)$, $log(F2/F1)$, and $log(F3/F2)$.

These analyses, while by no means exhaustive, provide evidence that spatial frequency relations may well play a part in the differences between DLs related to the reference formant patterns and should be further explored. However, we also continue to find, at this point, unexplainable differences related to the axis of movement.

Recall that three distinct patterns of spectral change are associated with movement away from reference points and that each pattern is related to one of the x', y', or z' axes.

Movement parallel to x' results in equal-percentage parallel shifts in $F1$ and $F2$ and may be thought of as reflecting Fant's (1960) distinction for grave/acute. The spectral patterns generated along these continua maintain a constant spatial frequency relation between $F1$ and $F2$, but vary the spectral location of this frequency. Movement parallel to y' results in equal-percentage opposing shifts in $F1$ and $F2$ and may be thought of as reflecting Fant's distinction for compact/diffuse. Spectral patterns generated along these continua vary the spatial frequency relation between $F1$ and $F2$, but maintain the spectral location of these frequencies. Movement parallel to z' results in unequal-percentage parallel shifts in $F1$, $F2$, and $F3$ and, while similar to x' movement, results in a more total shift of the spectrum. The spectral patterns generated along these continua maintain a constant spatial frequency relation between $F1$ and $F2$ and $F3$, but vary the spectral location of these frequencies.

Previous analyses have indicated that there is a small, but significant difference in discrimination between the spectral patterns associated with the x' and y' axes. We must assume for the present time that these differences are related to the differences in spatial frequency relations and spectral location noted above. While discrimination patterns of z' continua are somewhat similar to patterns seen for x' continua, the consideration of changes in $F3$ results in a distinct third pattern of discrimination requiring its own explanation. It should be the goal of future reseach to further elucidate the processes from which these differences in discrimination emanate.

Chapter 4

Final Comments and Implications for Future Research

It is suggested that the work presented in this dissertation be considered only a small example of the basic research yet required to understand the processes of speech perception. In addition, this work should also be considered exploratory in nature and, to that extent, the results preliminary. The results of the experiments presented here will hopefully open more doors to possible further investigation than they close as being definitive statements of fact. And that, I think, is as it should be.

Although the finding that abutting and non-overlapping target zones can be constructed based on synthetic vowel sounds lends validity to the target zone concept and the utilization of such zones in theories of speech perception, many issues remain unresolved and untested. How best to represent target zones, their boundaries, and the speech signal itself in order to accurately model the processes of human vowel perception will require considerable amounts of continued investigation. An immediate need for further research which has already been called for is the mapping of zone boundary areas at higher resoluton. Different methodological approaches, however, should be investigated to determine how judgements of tokens near category boundaries are influenced by subject task and other procedural variables. In addition, many simple extensions of Experiment I are also immediately implied. Similar mapping experiments with other fundamental frequency ($F0$) contours should im-

pact on current strategies for talker normalization and potentially related changes in zone boundaries. Considerable attention has been given to the potential differences in vowel perception between vowels in a consonantal context and vowels spoken in isolation (See Strange, 1989 and Neary, 1989 for discussions of these issues). Additional insights may be gained into these issues by comparing mappings of synthetic versions of vowels in and out of consonantal context.

Another dimension for continued investigation relates to the issue of non-vowel sounds located in the vowel space. As was discussed in the introduction to Chapter 2, portions of the unaccounted-for vowel space in target zone estimations based on natural speech productions (figures 1-9 and 1-10) may represent speech sounds other than vowels of American English or sounds not representative of speech at all. Miller and Hawks (1986) presented data which suggested that the initial portions of oral consonants and the voiced portions of fricatives may be best represented in areas of the space adjacent to vowel target zones. Additionally, it has been discussed that some subjects reported certain tokens were best identified as examples of front rounded vowels used in languages other than English. Since no effort was made in Experiment I to differentiate American English vowel sounds from other vowel or non-vowel sounds, some of the area within the synthetic-speech-based target zones may be misrepresented. Further research is required to resolve this issue.

The target zone concept within the *APS* format lends itself well for investigating aspects of perception and production for vowels of all languages. Some preliminary studies have already considered target zones for production data from Greek and German (Jongman, Fourakis, and Sereno, 1989) and perceptual data from one speaker of Greek (Fourakis and Hawks, 1990). In addition to comparative cross-language studies, these approaches may also prove valuable for investigating the effects of bilingualism and second-langauge acquisition on perception and production.

The utilization of approaches like those demonstrated in Experiment I for mapping perceptual responses need not be limited to vowels. Similar mapping studies for virtually any class of speech sounds is certainly plausible. Along with this plausibility comes all the possible variations already discussed for vowels. Consideration of these possible extensions suggests that the work presented here may well serve as only a small foundation on which

to build a sizeable body of new knowledge in speech science.

Of the number of findings related to the estimation of difference limens for multiple simultaneous formant changes of vowel-like sounds reported in Experiment II, perhaps the most important is that the *APS* vowel space, at least in terms of axis-parallel movement in the x', y', z' coordinate system, is not represented in equal perceptual dimensions. This finding impacts not only on considerations for target zone estimation and zone boundary representation, but also creates difficulties in implying any intuitive sense about the movement of the speech signal we are visualizing in these transformed dimensions. An acceptable condition would be to find that fixed distance movements parallel to the original x, y, z axes in *APS* do result in equal DLs, however, this too is probably not the case. Although percentages of formant change for fixed distance movements parallel to any of the x, y, z axes are equivalent, which formant or formants are allowed to vary is still axis dependent. Movement parallel to the x axis results in only $F3$ variation, while similar movement to y yields variations in all of the first three formants, and for z, variations in $F2$ and $F3$. Thus, no coordinate system currently utilized in the context of *APS* can be considered reflective of dimensions scaled in equal perceptual units for discrimination sensitivity. This was also the conclusion of Macmillan, et al. (1988), based on their attempts to account for differences in discrimination sensitivity along a vowel continuum through the examination of distances between the locations of continuum tokens in the *APS*.

Further investigative attention must be given to the implications of unequal perceptual scaling in *APS* in terms of how it may impact on practical and theoretical aspects of modeling human speech perception. Should equal perceptual scaling be required, an alternative coordinate system or further transformations of existing systems will be implied. Figure 4-1 demonstrates one possibility of a simple transformation. In this figure the synthetic-speech-based target zones for one z' plane are shown with the y' axis "warped" to reflect dimensions that are perceptually more equivalent.

A question of some import pertaining to perceptual considerations of the *APS* is what inferences can justifiably be drawn from the results of Experiment II which are reflective of human speech processing? The methodological considerations given to Experiment II were intended to force subjects into non-phonetic auditory processing modes, minimize the

possible effects of internal standards, and provide a relatively low-uncertainty environment for discriminating differences between two complex sounds. The issue can certainly be raised that such an experiment most probably does not reflect the processes used by listeners in everyday speech communication, and thus becomes another example of what humans "*can* do" and not what they "*do* do." This is, however, an issue which is ever present in speech research as well as other areas of science dealing with human abilities and capabilities and can never be completely dismissed. Many studies have been undertaken attempting to quantify the perceptual aspects of complex sound discrimination in both phonetic and non-phonetic contexts. The results of these studies vary as widely as the questions posed and the methodological procedures employed in them. Given that the true dimensions of the processes used in everyday speech perception have yet to be defined, the results of Experiment II potentially reflect the opposite end of the discrimination continuum. That is, the experiment attempted to demonstrate the maximum perceptual discrimination of vowel-like sounds resulting from a non-phonetic auditory processing mode within the *APS* context. If the results are representative of this description, we may at least rest assured that no smaller perceptual unit need be considered and that discrimination ability of vowels in natural speech communication modes should be based in larger perceptual units. The quantification of these units however, must await future research.

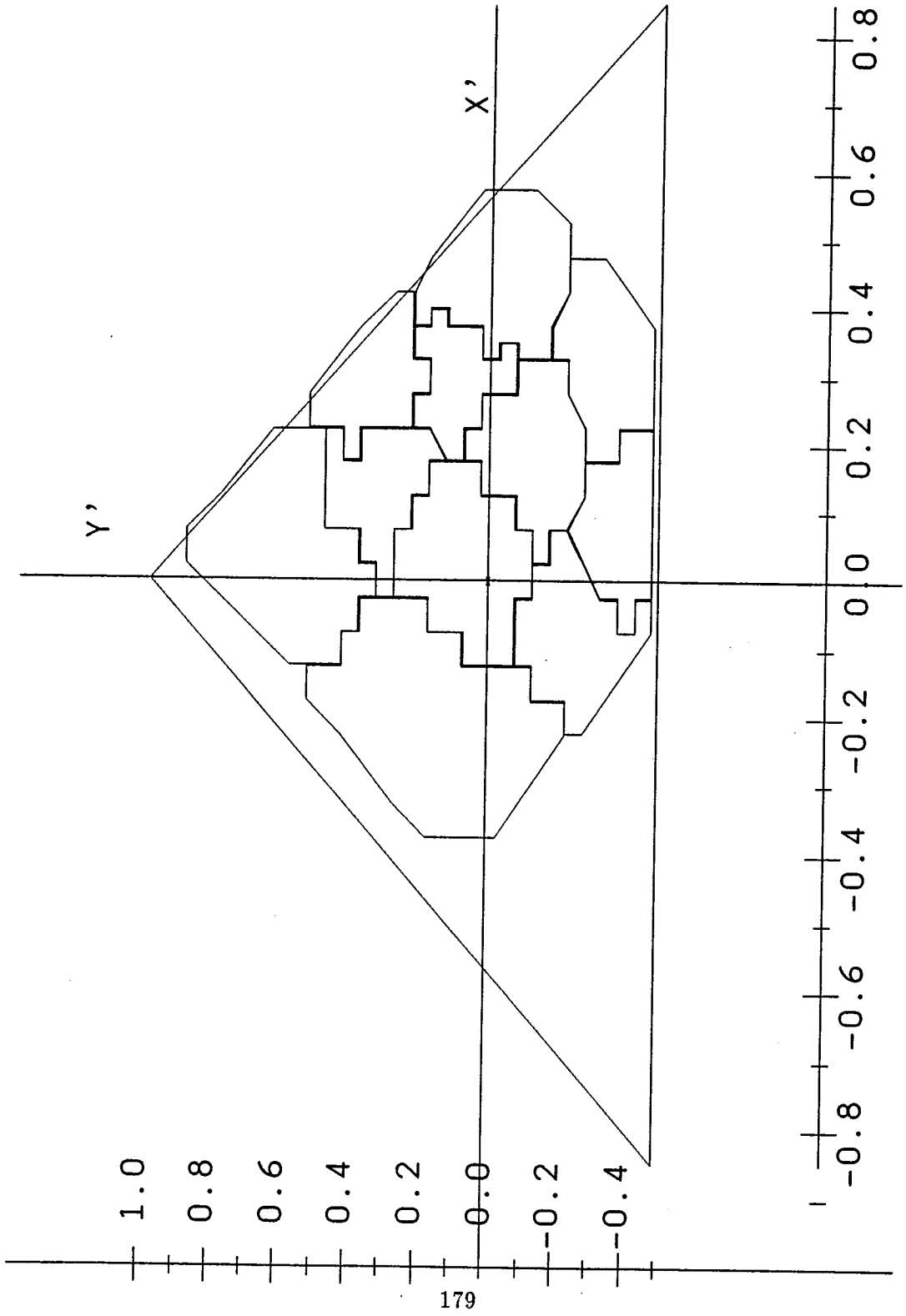
Additional motivation for further research can be spurred by other findings resulting from Experiment II. While the results found for discrimination of vowel-like sounds with simultaneously varying, multiple-formant changes may be generally accounted for by a relatively simple additive model where each formant contributes independent information to perception based on the discrimination of sounds varying in that formant only, the differences in discrimination found for patterns of parallel and opposing formant movement (i.e., axis orientation in the *APS*) and differences related to the formant patterns of reference points (i.e., location in the *APS*) are not so easily explained. Alternative methods of interpretation must be employed and investigated to understand these differences. For example, a previous discussion has indicated some plausibility for explanations based on the use of spatial frequency dimensions for the expression of changes in spectral patterns. Other models and metrics also exist for representing spectral changes as perceptual distances and

should be considered. In total, the results from Experiment II suggest a number of continuing avenues for future investigation into the realm of discrimination and perception of complex sounds.

Taken as a whole, the experiments presented here may not only both be applied and extended to further understanding of normal auditory processes, but also to a related sphere of research pertaining to the discrimination abilities and deficits of the hearing-impaired. While work in this area has been ongoing for many years, the need for a thorough understanding of auditory and speech processing in the impaired ear have never been greater. The advent of new technology surrounding cochlear implants and digital processing in hearing aids provides the potential for enhancing speech and other auditory signals in real time, a capability only dreamed of not many years ago. Natural extensions of the present experiments investigating differences in target zone boundary perception and discrimination of complex changes in vowel-like sounds between normal and impaired ears are justifiably called for, in that they should provide information beneficial to the further development of speech enhancement methods.

It is the hope of the author that, on the basis of the experiments presented here and the implications for further research stemming from them, the potential for target zone theory and the *APS* conceptual format to serve speech science as a research tool has been demonstrated. Additionally, it is hoped that this work may, in some way, prove to further the advancement of speech science, our knowledge of human perceptual abilities, and our general understanding of the communication process.

Figure 4-1: Locations of *SSB* target zones in APS $x'y'$ coordinates with axes modified to reflect approximately equal DL units.



Bibliography

Ainsworth, W.A. (**1972**). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am., **51**, 648-651.

Ainsworth, W.A. (**1975**). "Intrinsic and extrinsic factors in vowel judgements," in G. Fant and M. Tatham (eds.) *Auditory Analysis and Perception of Speech.* Academic Press, London, pp. 103-113.

Ainsworth, W.A., and Millar, J.B. (**1971**). "Methodology of experiments on the perception of synthesized vowels," Language and Speech, **14**, 201-212.

Assmann, P., Nearey, T., and Hogan, J. (**1982**). "Vowel identificaton: Orthographic, perceptual and acoustic aspects," J. Acoust. Soc. Am., **71**, 975-989.

Burdick, C.K., and Miller, J.D. (**1975**). "Speech perception by the chinchilla: discrimination of sustained /a/ and /i/," J. Acoust. Soc. Am., **58**, 415-427.

Carlson, R., Granström, B., and Klatt, D. (**1979**). "Vowel perception: The relative perceptual salience of selected acoustic manipulations," STL-QPSR, **3-4**, 73-83.

Clarke, F.R. (**1960**). "Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests," J. Acoust. Soc. Am., **32**, 35-46.

Cohen, J. (**1960**). "A coefficient of agreement for nominal scales," Ed. and Psych. Measurement, **XX**, No. 1, 37-46.

Disner, S.F. (**1980**). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am., **67**, 253-261.

Dunn, H.K. (**1961**). "Methods of measuring vowel formant bandwidths," J. Acoust. Soc. Am., **33**, 1737-1746.

Durlach, N.I., and Braida, L.D. (**1969**). "Intensity perception: I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am., **46**, 372-383.

Fairbanks, G., and Grubb, P. (**1961**). "A psychophysical investigation of vowel formants," J. Sp. Hear. Res., **4**, 203-219.

Fant, G. (**1973**). *Speech Sounds and Features.* MIT Press, Cambridge, MA., 227 p.

Fant, G. (**1972**). "Vocal tract wall effects, losses, and resonance bandwidths," STL-QPSR, **2-3**, 28-52.

Fant, G. (**1967**). "Auditory patterns of speech," in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 111-125.

Fant, G. (**1960**). *Acoustic Theory of Speech Production.* Mouton & Co., The Hague, The Netherlands.

Flanagan, J.L. (**1955**). "A difference limen for vowel formant frequency," J. Acoust. Soc. Am., **27**, 613-617.

Fourakis, M.S., and Hawks, J.W. (**1990**). "On the perceptual vowel space of Modern Greek," J. Acoust. Soc. Am., **87**, S159.

Fourakis, M.S., and Miller, J.D. (**1987**). "Measurement of vowels in isolation and in sonorant context," J. Acoust. Soc. Am., **81**, S17.

Fry, D.B., Abramson, A.S., Eimas, P.D., and Liberman, A.M., (**1962**). "The identification and discrimination of synthetic vowels," Lang. and Speech, **5**, 171-189.

Fujimura, O., and Lindqvist, J. (**1971**). "Sweep-tone measurements of vocal-tract characteristics," J. Acoust. Soc. Am., **49**, 541-558.

Fujisaki, H., and Kawashima, T. (**1968**). "The roles of pitch and the higher formants in the perception of vowels," IEEE Trans. Audio Electroacoust. **AU-16**, 73-77.

Greenberg, G.Z. (**1962**). "Cueing signals and frequency uncertainty in auditory detection," Tech. Rep. AF 19(628)-266 U.S.A.F.

Hawks, J.W. and Miller, J.D. (**1989**). "Perception of synthetic vowels: A comparison of several classification schemes," J. Acoust. Soc. Am., **86**, S78.

Holmes, J. (**1986**). "Normalization in vowel perception," in J. Perkell and D. KLatt (eds.) *Invariance and Variability in Speech Processes.* Lawrence Erlbaum, Hillsdale, NJ, pp. 346-357.

House, A.S., and Stevens, K.N. (**1958**). "Estimation of formant bandwidths from measurements of transient response of the vocal tract," J. Sp. Hear. Res., **1**, 309-315.

Jongman, A., Fourakis, M., and Sereno, J. (**1989**). "The acoustic vowel space of Modern Greek and German," Lang. and Speech, **32**, 221-248.

Kahn, D. (**1978**). "On the identifiability of isolated vowels," UCLA Working Papers in Phonetics, **41**, 26-31.

Kakusho, O., and Kato, K. (**1968**). "Just discriminable change and matching range of acoustic parameters of vowels," Acustica, **20**, 46-54.

Khazatsky, V. (**1985**). "Computational model of the effects of standards," Unpublished manuscript.

Klatt, D.H. (**1977**). "A cascade-parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis," J. Acoust. Soc. Am., **61**, S68.

Klatt, D.H. (**1979**). "Speech perception: A model of acoustic-phonetic analysis and lexical access," J. Phonetics, **7**, 279-312.

Klatt, D.H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., **67**, 971-995.

Klatt, D.H. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, 1278-1281.

Klatt, D.H. (**1987**). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., **82**, 737-757.

Klatt, D.H., and Klatt, L.C. (**1990**). "Analysis, synthesis and perception of voice quality variaitons among female and male talkers," J. Acoust. Soc.Am., **87**, 820-838.

Koenig, W. (**1949**). "A new frequency scale for acoustic measurements," Bell Labs Record, **27**, 299-301.

Kollmeier, B., Gilkey, R.H., and Sieben, U.K. (**1988**). "Adaptive staircase techniques in psychophysics: A comparison of human data and a mathematical model," J. Acoust. Soc. Am., **83**, 1852-1862.

Ladefoged, P. (**1982**). *A Course in Phonetics.* Harcourt Brace Jovanovich, Inc., New York, NY, 300 p.

Lee, W.A., and Shoup, J.E. (**1980**). "Specific contribution of the ARPA SUR project," in *Trends in Speech Recognition.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lehiste, I., and Peterson, G. (**1961**). "Transitions, glides, and diphthongs," J. Acoust. Soc. Am., **33**, 268-277.

Levitt, H. (**1970**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am., **49**, 65-69.

Liberman, A.M., Cooper, F.S., Shankweiler, D.S., and Studdert-Kennedy, M. (**1967**). "Perception of the Speech Code," Psychol. Rev., **74**, 431-461.

Liberman, A.M., and Mattingly, I.G. (**1985**). "The motor theory of speech perception revised," Cognition, **21**, 1-36.

Liljencrants J., and Lindblom, B. (**1972**). "Numerical simulation of vowel quality systems: The role of perceptual contrast," Language, **48**, 839-862.

Luce, R.D. (**1963**). "A threshold theory for simple detection experiments," Psych. Rev., **70**, 61-79.

Macmillan, N.A., Braida, L.D., Goldberg, R.F., and Khazatsky, V. (**1986**). "Central and peripheral processes in the perception of speech and nonspeech sounds," NATO Conference, Utrecht, The Netherlands.

Macmillan, N.A., Goldberg, R.F., and Braida, L.D. (**1988**). "Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am., **84**, 1262-1280.

Macmillan, N.A., Kaplan, H.L., and Creelman, C.D., (**1977**), "The psychophysics of categorical perception," Psych. Rev., **84**, 452-471.

McKay, D.M. (**1956**). "The Epistemological Problems for Automata," in C.E. Shannon and J. McCatthy (eds.) *Automata Studies.* Princeton University Press, Princeton, NJ.

Mermelstein, P. (**1978**). "Difference limens for formant frequencies of steady-state and consonant-bound vowels," J. Acoust. Soc. Am., **63**, 572-580.

Millar, J.B., and Ainsworth, W.A. (**1972**). "Identification of synthetic isolated vowels and vowels in H-D context," Acustica, **27**, 278-282.

Miller, J.D. (**1980**). "Estimation of formant bandwidths for vowels," Unpublished.

Miller, J.D. (**1984**). "Auditory processing of the acoustic patterns of speech," Arch. Otolaryngol., **110**, 154-159.

Miller, J.D. (**1987b**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am., **81**, S16.

Miller, J.D. (**1987c**). "Classifications of vowel productions by means of perceptual target zones: A response to Ladefoged and Studdert-Kennedy," J. Acoust. Soc. Am., **82**, S82.

Miller, J.D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am., **85**, 2114-2121.

Miller, J.D., and Hawks, J.W. (**1986**). "Spectral envelopes and perceptual target zones for consonants and vowels: Preliminary estimates," J. Acoust. Soc. Am., **79**, S66.

Miller, J.D., and Hawks, J.W. (**1989**). "Target zones for synthetic vowels," J. Acoust. Soc. Am., **85**, S66.

Miller, R.L. (**1953**). "Auditory tests with synthetic vowels," J. Acoust. Soc. Am., **18**, 114-121.

Morton, J., and Broadbent, D.E. (**1967**). "Passive versus active recognition models or is your homunculus really necessary?" in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 103-110.

Neary, T.M. (**1977**). "Phonetic feature systems for vowels," Ph.D. Thesis, Univ. of Connecticut. Reproduced by Indiana University Linguistics Club, 1978.

Neary, T.M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am., **85**, 2088-2113.

Nord, L., and Sventelius, E. (**1979**). "Analysis and prediction of difference limen data for formant frequencies," STL-QPSR, **3-4**, 60-72.

Peterson, G.E., and Barney, H.L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am., **24**, 175-184.

Peterson, G.E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am., **32**, 693-703.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech, J. Sp. Hear. Res., **29**, 434-446.

Pike, K. (**1947**). "On the phonemic status of English diphthongs," Language, **23**, 151-159.

Pisoni, D.B. (**1971**). "On the Nature of categorical perception of speech sounds." Ph.D. Thesis. University of Michigan.

Pisoni, D.B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Perception and Psychophysics, , **13**, 253-260.

Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones, " in R. Plomp and G.F. Smoorenburg (Eds.) *Frequency Analysis and Periodicity Detection in Hearing.* Sijthoff, Leiden, The Netherlands, pp. 397-414.

Pollack, I., and Decker, L.R. (**1958**). "Confidence ratings, message reception and the receiver operating characteristic," J. Acoust. Soc. Am., **30**, 286-292.

Pols, L.C.W., Van Der Kamp L. J. Th., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am., **46**, 458-467.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (**1973**). "Frequency analysis of Dutch vowels from 50 male speakers," J. Acoust. Soc. Am., **53**, 1093-1101.

Potter, R.K., and Peterson, G.E. (**1948**). "The representation of vowels and their movements," J. Acoust. Soc. Am., **20**, 528-535.

Repp, B.H., Healy, A.F., and Crowder, R.G. (**1979**). "Categories and context in the perception of isolated steady-state vowels," J. Exp. Psychol.; Human Perception and Performance, **5**, 129-145.

Robinson, D.E., and Watson, C.S. (**1972**). "Psychophysical methods in modern psychoacoustics," In J.V. Tobias (Ed.) *Foundations of Modern Auditory Theory.* Academic Press, New York, NY, pp. 101-131.

Sachs, R.M. (**1969**). "Vowel identification and discrimination in isolation vs. word context," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M.I.T., 220-229.

Scholes, R.J. (**1967**). "Categorial responses to synthetic vocalic stimuli by speakers of various languages," Lang. and Speech, **10**, 252-282.

Stevens, K.N. (**1968**). "On the relations between speech movements and speech perception," Zeitschr. f. Phon., **21**, 102.

Stevens, K.N., and Halle, M. (**1967**). "Remarks on analysis by synthesis and distinctive features," in W. Wathen-Dunn (ed.) *Models for the Perception of Speech and Visual Form.* MIT Press, Cambridge, MA. pp. 88-102.

Stevens, K.N., Liberman, A.M., Studdert-Kennedy, M., and Öhman, S.E.G. (**1969**). "Crosslanguage study of vowel perception," Lang. and Speech, **12**, 1-23.

Strange, W. (**1989**). "Evolving theories of vowel perception," J. Acoust. Soc. Am., **85**, 2081-2087.

Strange, W., Edman, T.R., and Jenkins, J.J. (**1979**). "Acoustic and phonological factors in vowel identification," J. Exp. Psych.: Human Perception and Performance, **5**, 643-656.

Strange, W., Verbrugge, R., Shankweiler, D., and Edman, T. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am., **60**, 213-224.

Syrdal, A., and Gopal, H. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am., **79**, 1086-1100.

Tatsuoka, M. (**1970**). *Selected Topics in Advanced Statistics: An Elementary Approach Pt. 6: Discrimination Analysis* (Institute for Personality and Ability Testing, Champaign, IL).

Taylor, M.M., and Creelman, C.D. (**1967**). "PEST: Efficient estimates on probability functions," J. Acoust. Soc. Am., **41**, 782-787.

Traunmüller, H. (**1981**). "Perceptual dimensions of openness in vowels," J. Acoust. Soc. Am., **69**, 1465-1475.

Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am., **68**, 1523-1525.

Appendix A

Formant location in the *APS*

This section will demonstrate and discuss how certain formant patterns manifest themselves in x', y', z' space and will be limited to cases where the sensory reference (SR) remains fixed. Recall from section 1.2.1 that the auditory-perceptual space (APS) is defined in three dimensions, x, y, and z, where

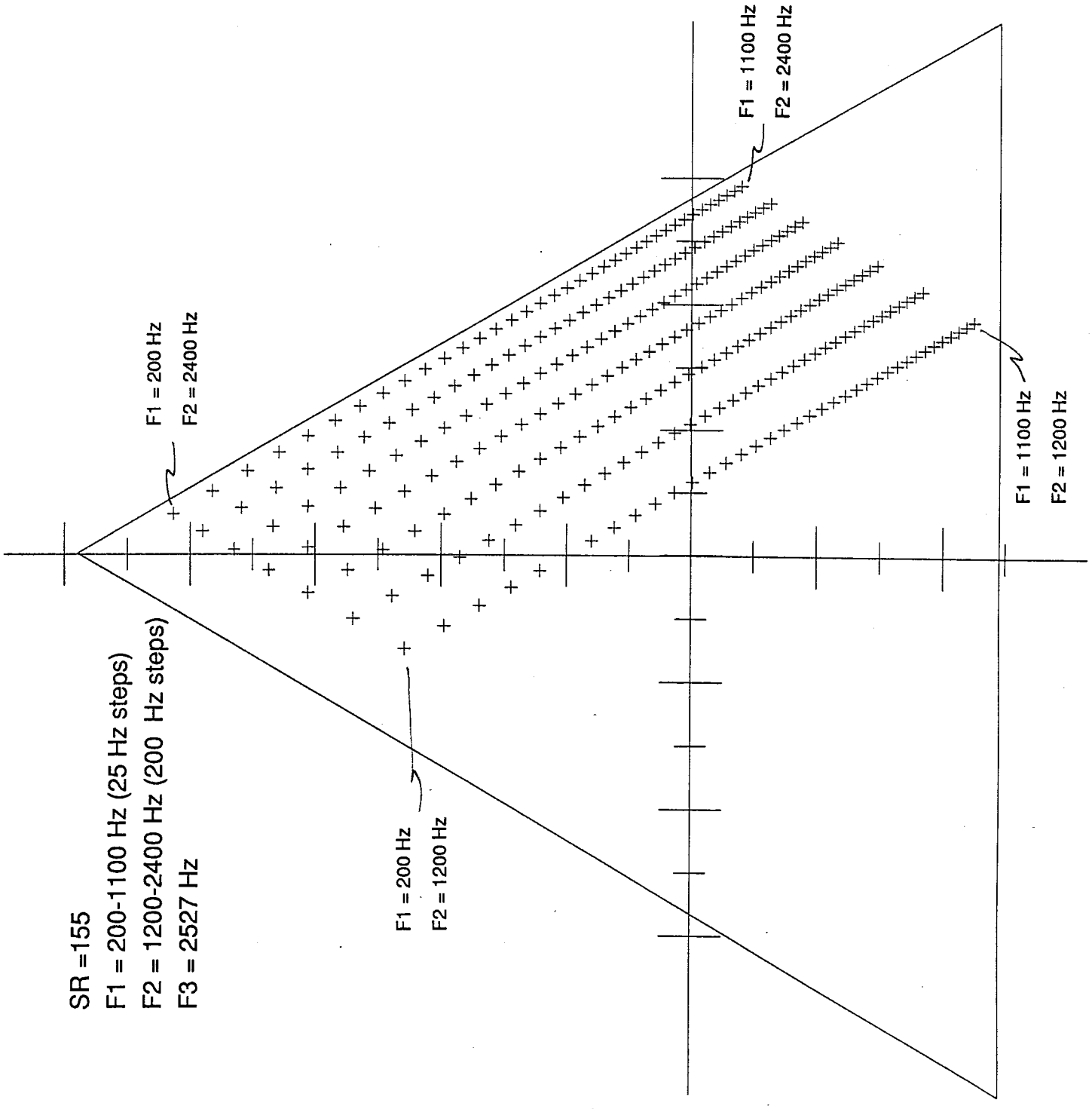
$$\begin{aligned} x &= log(SF3/SF2), \\ y &= log(SF1/SR), \quad and \\ z &= log(SF2/SF1). \end{aligned} \tag{A.1}$$

These coordinates are often transformed for visual simplification by rotation the APS axes, yielding a new set of coordinates, x', y', z', where,

$$\begin{aligned} x' &= .70711(y - x), \\ y' &= .8162(z) - .4081(x + y), and \\ z' &= .5772(x + y + z). \end{aligned} \tag{A.2}$$

Figure A-1 illustrates the location of points generated in $x'y'$-space when $F2$ and $F3$ are fixed and only $F1$ is allowed to vary. The figure shows seven examples of this configuration with different values of $F2$. Such configurations might be approximated in natural speech as formant movements from /AE/ to /IY/ or /AA/ to /UW/. As $F1$ increases, points associated with a given $F2$ move down and to the right creating a line of points which lies at a 60° angle relative to the x' axis. These lines represent constant values of $F2$. All points lie in

Figure A-1: Location of seven continua generated in $x'y'$ space with fixed values of $F2$ and $F3$ with $F1$ allowed to vary.



a single z' plane ($z' = 0.70$) and, since $F1$ is incremented in linear steps, are logarithmically spaced. Figure A-2 demonstrates somewhat the opposite angular effect. Here $F1$ and $F3$ are fixed and only $F2$ allowed to vary. Once again seven examples are illustrated with different values of $F1$. Similar configurations to these might be approximated in natural speech as formant movements from /UW/ to /IY/ or /IH/, or /AO/ to /AE/ or /EH/. All points still lie in a single z' plane and the points associated with a given $F1$ form a line now at a 120° angle to the x' axis.

If $F1$ and $F2$ are held constant and $F3$ is allowed to vary, the patterns seen in Figure A-3 emerge. These patterns might be similar to paths seen when the back vowels /UW/ or /AO/ are roticized, as when they precede /R/ in natural speech. Two groups of data points are presented with one group representing a fixed value for $F1$ with several values of $F2$ and another group representing a fixed value of $F2$ with several values of $F1$. Irrespective of the groupings, a line of points lying at a 150° angle relative to the x' axis is generated for each pair of $F1$ and $F2$ values. Figure A-4 provides a "side view" of the same data points in $y'z'$-coordinate space. The lines of points move "forward" along z' (left to right in this view) at a 33° angle to the z' axis as $F3$ increases. Thus the "lines" of points for pairs of $F1$ and $F2$ seen in Figure A-3 are somewhat deceiving in that the movement is occuring along the z' axis, and therefore, for any given value of z' only a single point would be present for each formant pair.

If $F1$ and $F2$ maintain a constant ratio with each other (with $F3$ fixed), approximating formant movements from /UW/ to /AE/ in natural speech, points of like ratios form lines falling parallel to the x' axis. This is illustrated in Figure A-5 which shows eight horizontal lines of points where $log(F1/F2) = 0.10$ to 0.38 in 0.04 steps. However, these log ratios may also be expressed as simple formant ratios where $F2/F1 = 1.26$ to 2.4. Once again all points lie stationary in one z' plane.

Figure A-6 demonstrates the effect of formants moving toward or away from each other. In this figure, the lowest point represents a merged $F1$ and $F2$, similar to /AA/, with data points moving upward along the y' axis as $F1$ and $F2$ move away from each other, to an /IH/ or /IY/ configuration. Again, with $F3$ fixed, there is no movement along the z' axis. For lines of data points to lie parallel to the y' axis and have no movement along z', the

Figure A-2: Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F3$ with $F2$ allowed to vary.

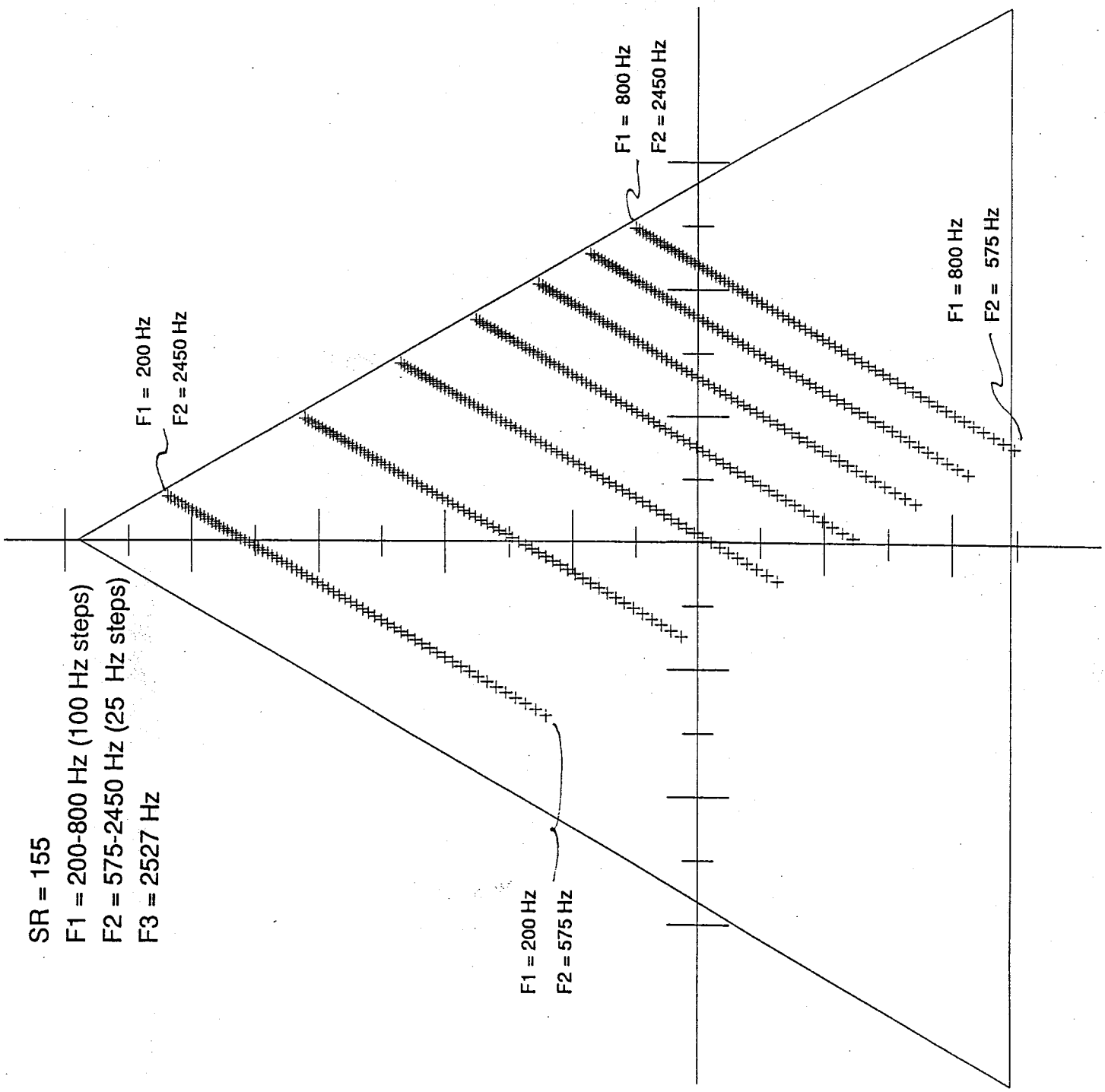


Figure A-3: Location of seven continua generated in $x'y'$ space with fixed values of $F1$ and $F2$ with $F3$ allowed to vary. Crosses indicate continua with a fixed $F1$ and $F2$ changing with each continuum. Squares indicate continua with a fixed $F2$ and $F1$ changing with each continuum.

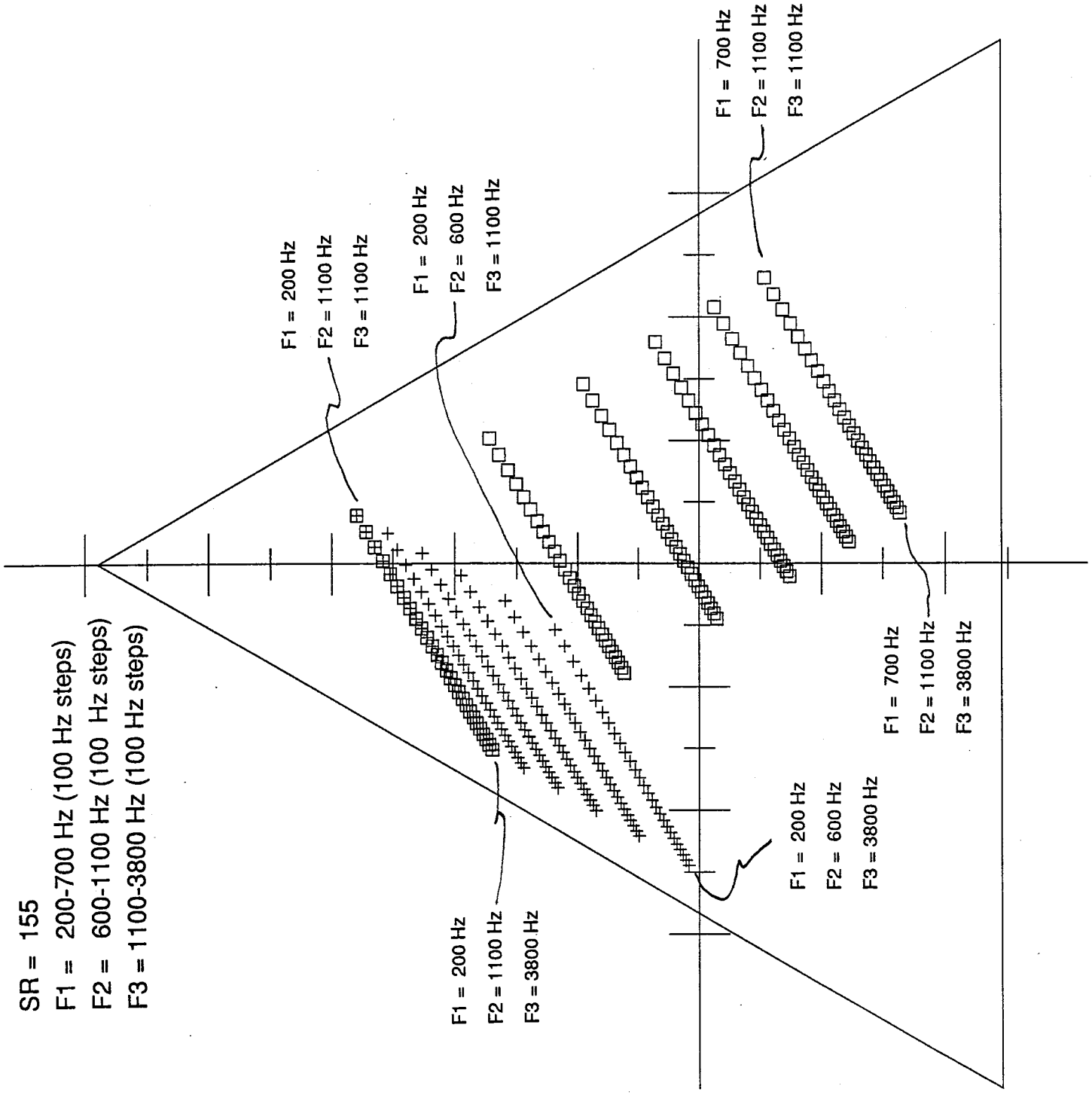


Figure A-4: "Side" view in $y'z'$ space of continua from Figure A-3.

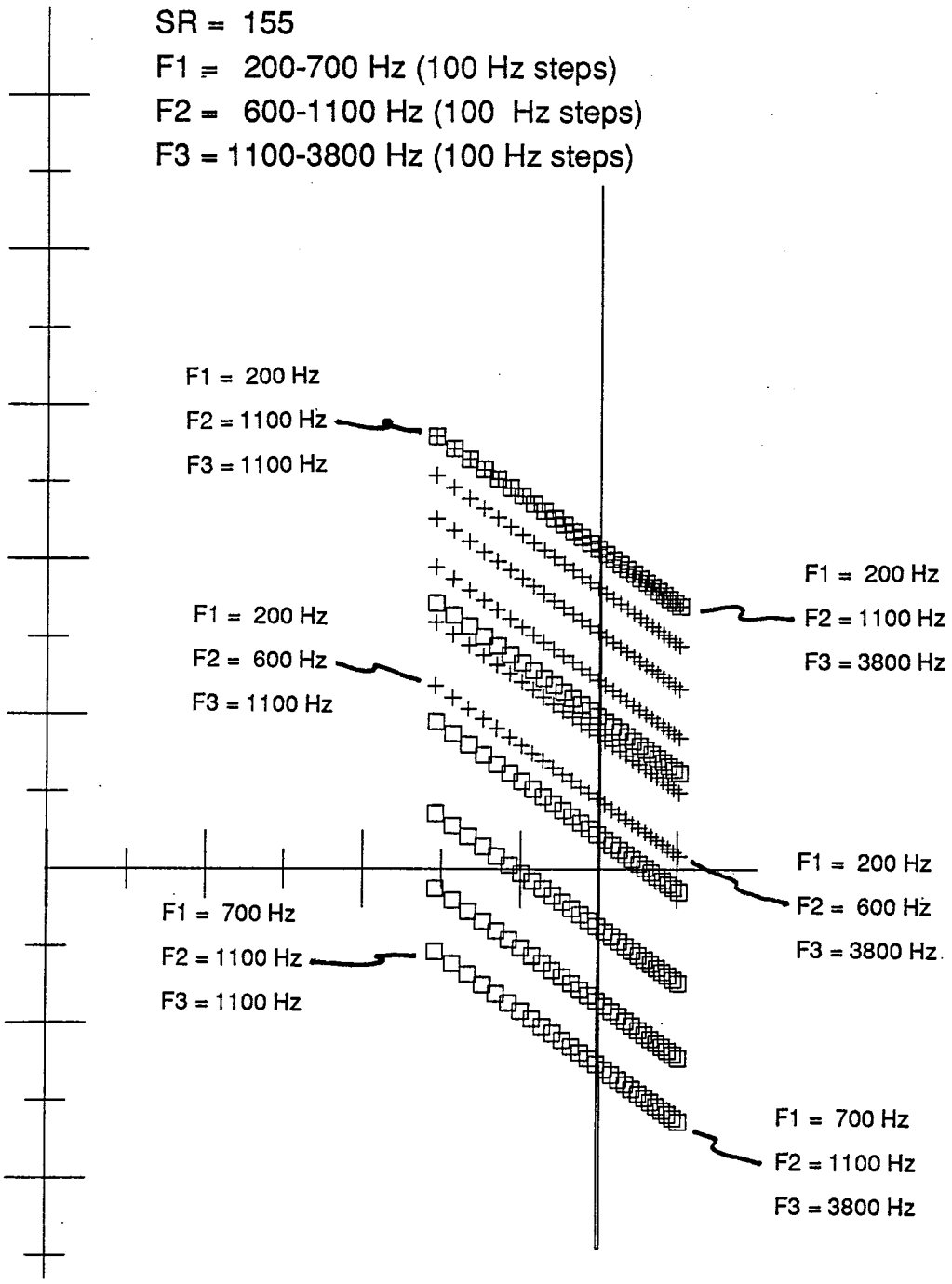


Figure A-5: Location of eight continua generated in $x'y'$ space with a fixed $F3$ and $F1$ and $F2$ maintained in constant ratios.

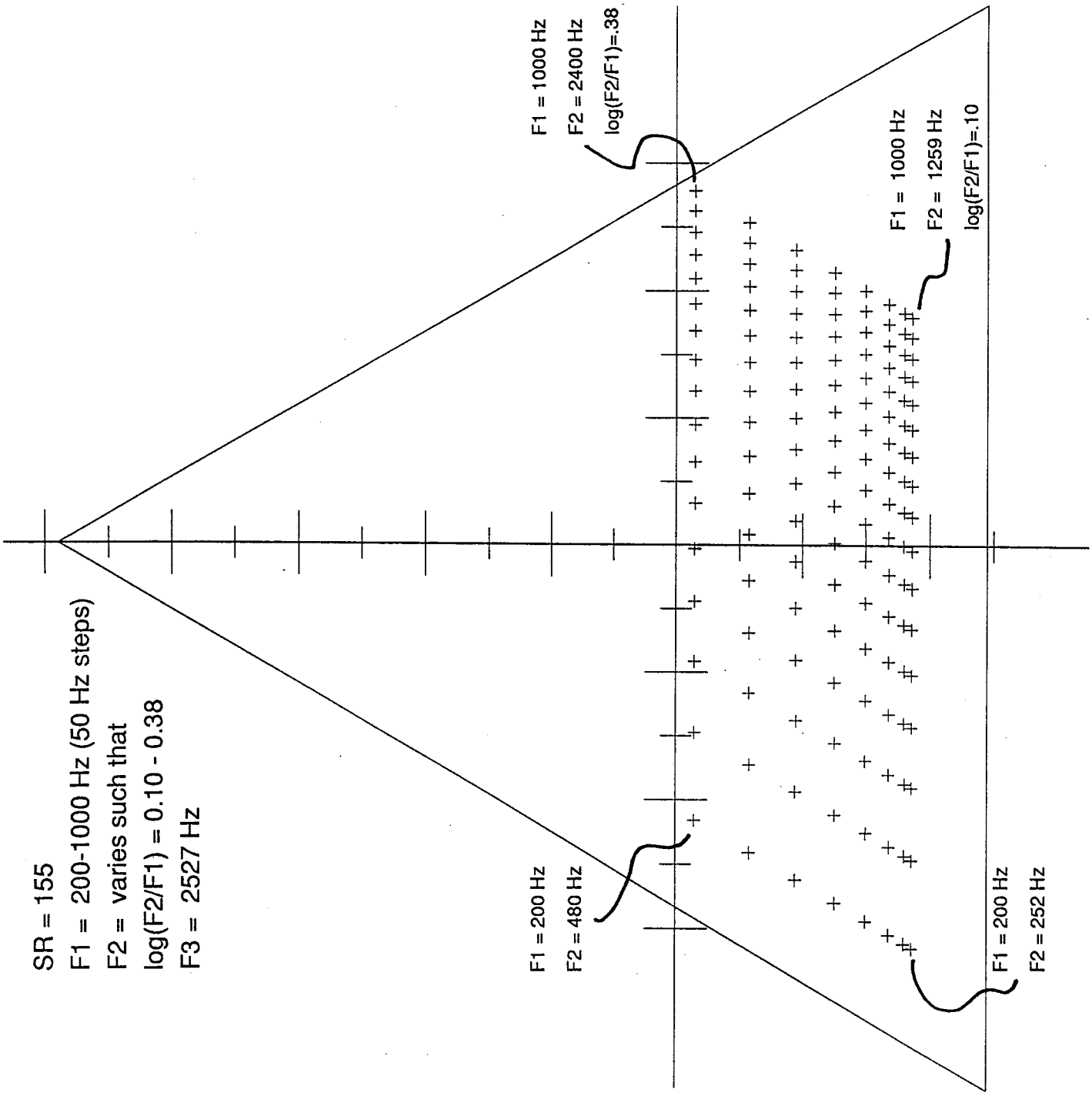
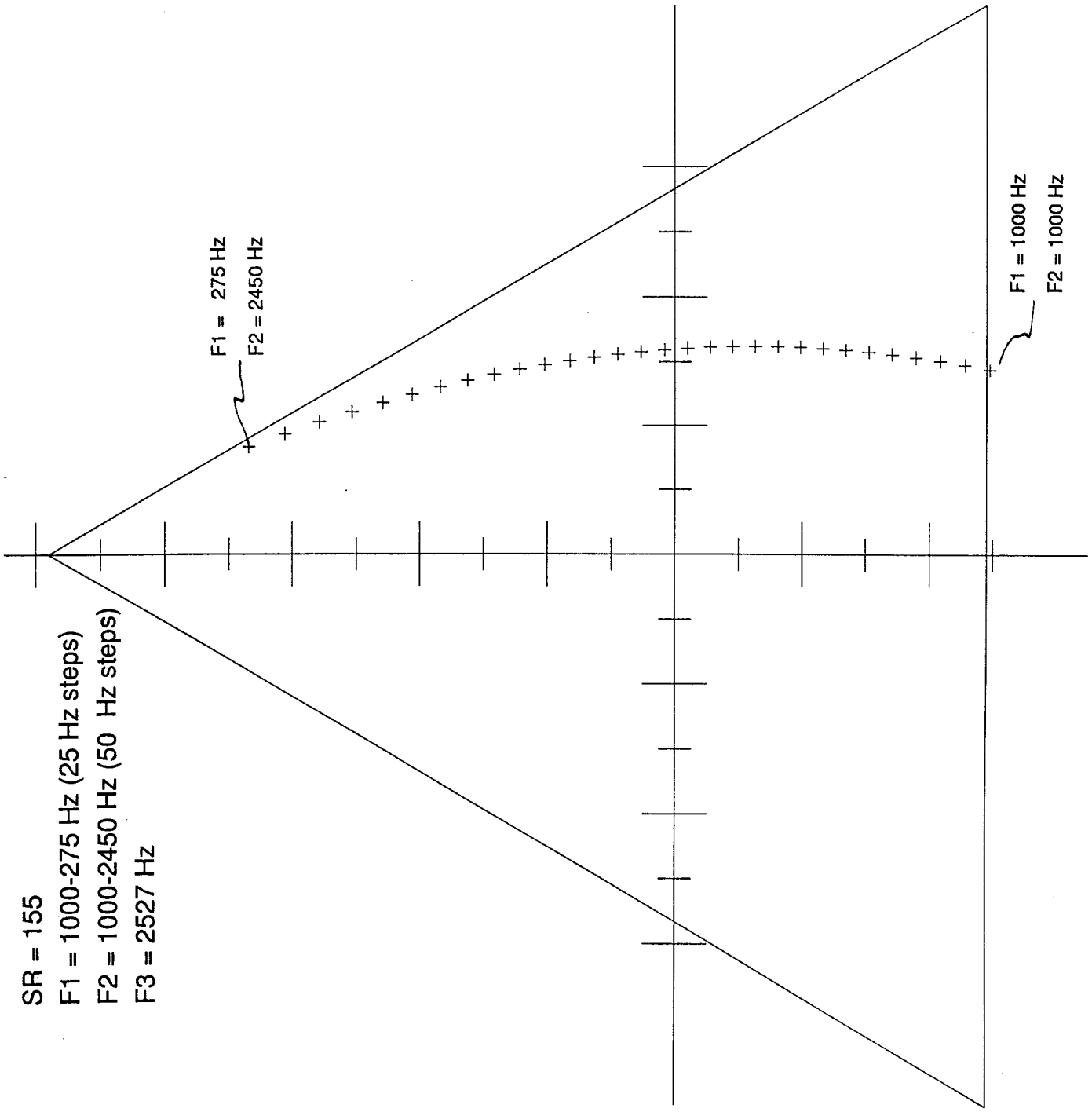


Figure A-6: Location of continuum generated in $x'y'$ space with $F3$ fixed and increasingly greater separation in $F1$ and $F2$.



values of $F1$ and $F2$ must be manipulated such that

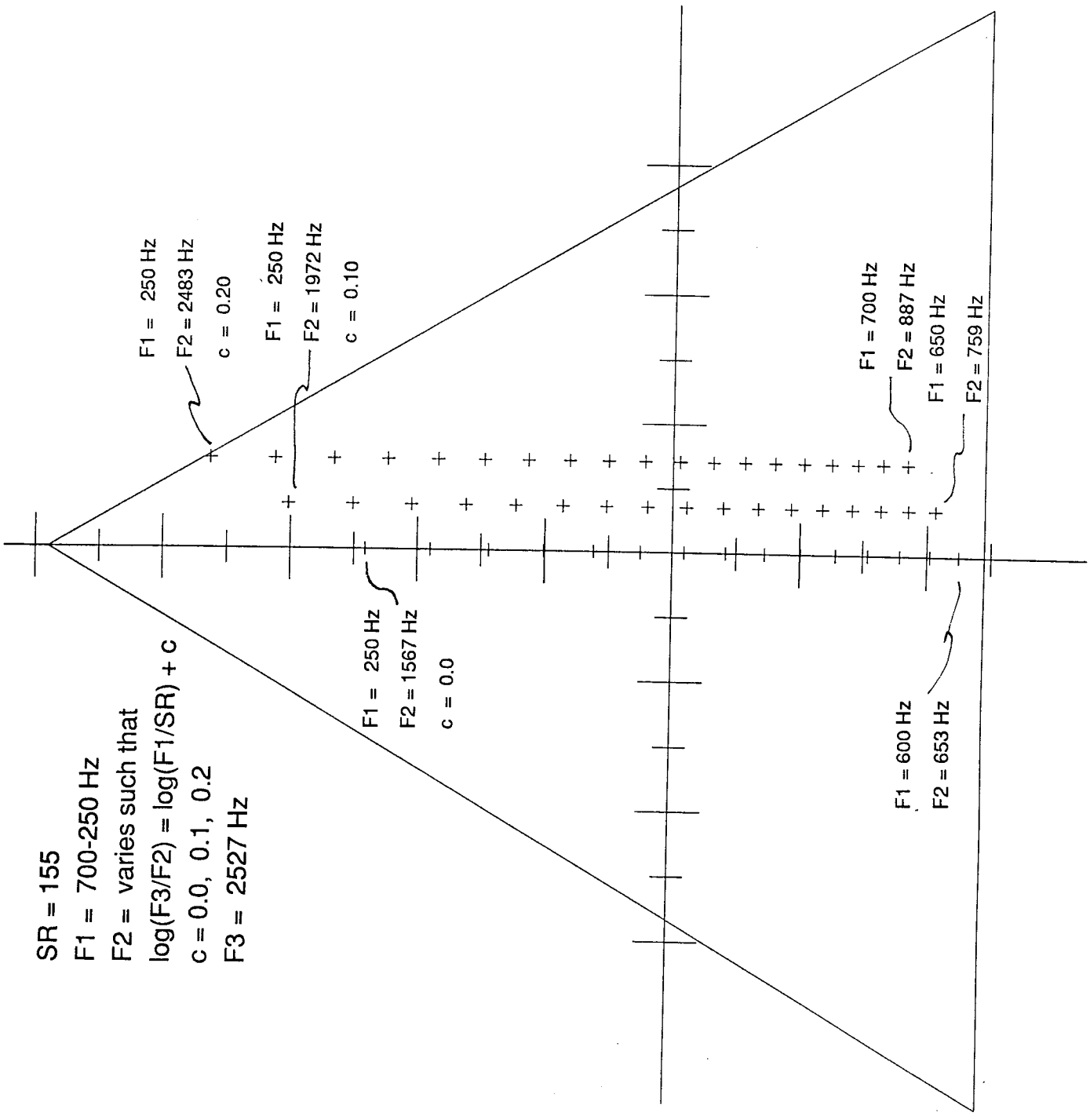
$$log(F3/F2) = log(F1/SR) + c, \tag{A.3}$$

where $F3$, SR, and c are constants. This is illustrated in Figure A-7 with each of the three vertical rows of points reflecting a different value of c. Although this parallel movement along the y' axis appears complicated, the more general aspect of movement along this axis is that $F1$ and $F2$ are moving in opposite directions.

In summary, points generated with constant values of $F1$ fall along 120° lines relative to the x' axis with $F1$ increasing top-left to bottom-right in the $x'y'$ view. Points generated with constant values of $F2$ fall along 60° lines relative to the x' axis with $F2$ increasing bottom-left to top-right in the $x'y'$ view. Points genreated with constant values of $F3$ fall along lines which are perpendicular planes to the z' axis with $F3$ increasing "back to front" in the $x' - y'$ view. When both $F1$ and $F2$ are held constant, representative points fall along lines appearing to fall at a 150° angle to the x' axis in the $x'y'$ view, but are actually angular lines along the z' axis. When $F1$ and $F2$ both increase or decrease together, representative points fall along lines more parallel to the x' axis, while points representing values of $F1$ and $F2$ moving toward or away from each other fall along lines more parallel to y'.

The three-dimensional metric utilized by the *APS* provides a valuable and innovative method of visualizing changes in complex sounds. Although formant change in natural speech does not behave as precisely or as simply as those shown in the preceding figures, it is hoped that this section has provided a better understanding of how the locations and changes in speech sounds may be interpreted with such a metric.

Figure A-7: Location of three continua generated in $x'y'$ space parallel to the y' axis. SR, $F3$, and a constant c are fixed.



Appendix B

Synthesis Parameter Specifications

All parameters specified for the synthesis of stimuli are shown in Table B.1 as they appear in the synthesizer program. The columns under "SYM" indicate show the abbreviations for the parameters with the "v/C" columns indicating whether that parameter is of a variable or constant specification. The columns under "VAL" indicate values used for token synthesis with the exception of parameters F1, F2, F3, b1, b2, b3, g0 (overall gain) which could vary with each token, and parameters f0 and av, which are identical for each token, vary over time. The variations over time for these two parameters are shown in Table B.2. For those unfamiliar with the Klatt synthesis program, definitions for most variables can be found in Klatt, 1980; 1987 and Klatt and Klatt, 1990.

Bandwidths in Hertz by formant frequency based on data from Miller(1980) are shown in Table B.3. These values were used for the specification of bandwidths for $F1$, $F2$, and $F3$ for all synthetic tokens generated for both experiments. Formant frequencies falling between values specified in the table were assigned the lower bandwidth value. See Section 2.2.1 for additional details on formant bandwidth calculation.

Table B.1: Synthesis parameter specifications.

SYM	v/C	MIN	VAL	MAX	SYM	v/C	MIN	VAL	MAX
sr	C	5000	10000	20000	nf	C	1	4	8
du	C	30	400	5000	ss	C	1	2	2
ui	C	1	5	20	rs	C	1	1	99999
f0	v	0	1000	5000	av	v	0	60	80
F1	v	180	270	1300	b1	v	30	49	1000
F2	v	550	2290	3000	b2	v	40	105	1000
F3	v	1139	1139	4800	b3	v	60	152	1000
F4	v	2400	4000	4990	b4	v	100	500	1000
F5	v	3000	4900	4990	b5	v	100	1000	1500
f6	v	3000	4990	4990	b6	v	100	500	4000
fz	v	180	280	800	bz	v	40	90	1000
fp	v	180	280	500	bp	v	40	90	1000
ah	v	0	0	80	oq	v	10	50	80
at	v	0	0	80	tl	v	0	0	34
af	v	0	0	80	sk	v	0	0	100
a1	v	0	0	80	p1	v	30	80	1000
a2	v	0	0	80	p2	v	40	200	1000
a3	v	0	0	80	p3	v	60	350	1000
a4	v	0	0	80	p4	v	100	500	1000
a5	v	0	0	80	p5	v	100	600	1500
a6	v	0	0	80	p6	v	100	800	4000
an	v	0	0	80	ab	v	0	0	80
ap	v	0	0	80	os	C	0	0	20
g0	v	0	68	80	dF	v	0	0	100
db	v	0	0	400					

Table B.2: Time-varying synthesis parameter specifications for F0 (x10) and amplitude.

Time	F0	AV	Time	F0	AV	Time	F0	AV
0	1140	1	135	1320	55	270	1205	55
5	1156	6	140	1320	55	275	1196	55
10	1172	12	145	1320	55	280	1188	55
15	1189	17	150	1320	55	285	1180	55
20	1205	23	155	1320	55	290	1172	55
25	1221	28	160	1320	55	295	1164	55
30	1238	33	165	1320	55	300	1155	55
35	1254	39	170	1320	55	305	1147	55
40	1270	44	175	1320	55	310	1139	55
45	1287	50	180	1320	55	315	1131	55
50	1303	55	185	1320	55	320	1123	55
55	1320	55	190	1320	55	325	1114	55
60	1320	55	195	1320	55	330	1106	55
65	1320	55	200	1320	55	335	1098	55
70	1320	55	205	1311	55	340	1090	55
75	1320	55	210	1303	55	345	1082	55
80	1320	55	215	1295	55	350	1073	54
85	1320	55	220	1287	55	355	1065	53
90	1320	55	225	1278	55	360	1057	52
95	1320	55	230	1270	55	365	1049	50
100	1320	55	235	1262	55	370	1041	48
105	1320	55	240	1254	55	375	1032	44
110	1320	55	245	1246	55	380	1024	39
115	1320	55	250	1237	55	385	1016	31
120	1320	55	255	1229	55	390	1008	19
125	1320	55	260	1221	55	395	1000	1
130	1320	55	265	1213	55			
135	1320	55	270	1205	55			

Table B.3: Formant bandwidths (BW) by formant frequency (Frmt) in Hertz utilized for all synthetic tokens in all experiments.

Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW
133	71	1233	63	2238	103	2887	143	3385	183	3799	223
134	70	1267	64	2257	104	2901	144	3396	184	3808	224
136	69	1300	65	2276	105	2915	145	3407	185	3818	225
138	68	1333	66	2295	106	2929	146	3418	186	3827	226
140	67	1365	67	2313	107	2942	147	3429	187	3837	227
143	66	1396	68	2332	108	2956	148	3440	188	3846	228
145	65	1427	69	2350	109	2969	149	3451	189	3856	229
148	64	1457	70	2368	110	2983	150	3462	190	3865	230
151	63	1487	71	2386	111	2996	151	3473	191	3874	231
154	62	1516	72	2404	112	3009	152	3484	192	3884	232
158	61	1545	73	2421	113	3022	153	3495	193	3893	233
161	60	1573	74	2439	114	3035	154	3505	194	3902	234
166	59	1600	75	2456	115	3048	155	3516	195	3911	235
170	58	1628	76	2473	116	3061	156	3527	196	3921	236
175	57	1654	77	2490	117	3074	157	3537	197	3930	237
181	56	1681	78	2507	118	3087	158	3548	198	3939	238
188	55	1707	79	2523	119	3100	159	3558	199	3948	239
195	54	1732	80	2540	120	3112	160	3569	200	3957	240
203	53	1757	81	2556	121	3125	161	3579	201	3966	241
213	52	1782	82	2573	122	3137	162	3589	202	3975	242
225	51	1807	83	2589	123	3150	163	3600	203	3984	243
240	50	1831	84	2605	124	3162	164	3610	204	3993	244
259	49	1855	85	2621	125	3174	165	3620	205	4002	245
286	48	1878	86	2636	126	3186	166	3631	206	4011	246
332	47	1902	87	2652	127	3199	167	3641	207	4020	247
512	48	1925	88	2667	128	3211	168	3651	208	4029	248
600	49	1947	89	2683	129	3223	169	3661	209	4038	249
668	50	1970	90	2698	130	3235	170	3671	210	4047	250
727	51	1992	91	2713	131	3247	171	3681	211	4056	251
780	52	2013	92	2728	132	3258	172	3691	212	4064	252
830	53	2035	93	2743	133	3270	173	3701	213	4073	253
878	54	2056	94	2758	134	3282	174	3711	214	4082	254
923	55	2077	95	2773	135	3294	175	3721	215	4091	255
966	56	2098	96	2788	136	3305	176	3731	216	4099	256
1008	57	2119	97	2802	137	3317	177	3740	217	4108	257
1048	58	2139	98	2817	138	3328	178	3750	218	4116	258
1087	59	2159	99	2831	139	3340	179	3760	219	4125	259
1125	60	2179	100	2845	140	3351	180	3770	220	4134	260
1162	61	2199	101	2859	141	3362	181	3779	221	4142	261
1198	62	2219	102	2873	142	3374	182	3789	222	4151	262

Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW	Frmt	BW
4159	263	4317	282	4468	301	4612	320	4751	339	4885	358
4168	264	4325	283	4476	302	4620	321	4758	340	4892	359
4176	265	4333	284	4483	303	4627	322	4766	341	4899	360
4185	266	4341	285	4491	304	4635	323	4773	342	4905	361
4193	267	4350	286	4499	305	4642	324	4780	343	4912	362
4202	268	4358	287	4507	306	4650	325	4787	344	4919	363
4210	269	4366	288	4514	307	4657	326	4794	345	4926	364
4218	270	4374	289	4522	308	4664	327	4801	346	4933	365
4227	271	4382	290	4530	309	4672	328	4808	347	4940	366
4235	272	4389	291	4537	310	4679	329	4815	348	4946	367
4243	273	4397	292	4545	311	4686	330	4822	349	4953	368
4252	274	4405	293	4552	312	4694	331	4829	350	4960	369
4260	275	4413	294	4560	313	4701	332	4836	351	4967	370
4268	276	4421	295	4568	314	4708	333	4843	352	4973	371
4276	277	4429	296	4575	315	4715	334	4850	353	4980	372
4285	278	4437	297	4583	316	4723	335	4857	354	4987	373
4293	279	4445	298	4590	317	4730	336	4864	355	4993	374
4301	280	4452	299	4598	318	4737	337	4871	356	5000	375
4309	281	4460	300	4605	319	4744	338	4878	357		
4317	282	4468	301	4612	320	4751	339	4885	358		

Appendix C

Spectral Envelopes for Experiment II reference tokens

Table C.1 lists the values of $F1$, $F2$, and $F3$ for the reference tokens used in Experiment II. The figures that follow represent the spectral envelopes of the 17 reference point tokens utilized in Experiment II. The spectral envelopes for the center reference points can be considered as representive of tokens used in Experiment I as well. These envelopes are the results of 256-point FFTs centered at the 120 msec point of each token and based on the LPC coefficients generated from analyses with the ILS software package. The LPC analyses computed 12 coefficients based on a 25.6 msec hamming window shifted in 1 msec steps along the pre-emphasized signal. The vertical lines in the spectral envelopes represent points of measurement and not harmonic content. The figures are labeled in a fashion similar to that used in Figures 3-2 and 3-5.

Table C.1: Formant ($F1$, $F2$, $F3$) values for the 17 reference points used in Experiment II.

Reference	$F1$	$F2$	$F3$
IY	247	2275	3086
IH	411	1830	2528
EH	583	1785	2528
AE	828	1740	2528
AA	921	1329	2528
AH	627	1199	2528
AO	643	844	2528
UH	388	1188	2528
UW	227	764	2528
ER	401	1213	1390
IYIH	316	2089	2528
IHEH	475	1925	2528
EHAE	652	2072	2528
AEAH	739	1553	2528
AHAA	663	1133	2528
AHUH	459	1143	2528
UHUW	329	1046	2528

Figure C-1: Spectral envelope derived from FFT of [IY] reference token.

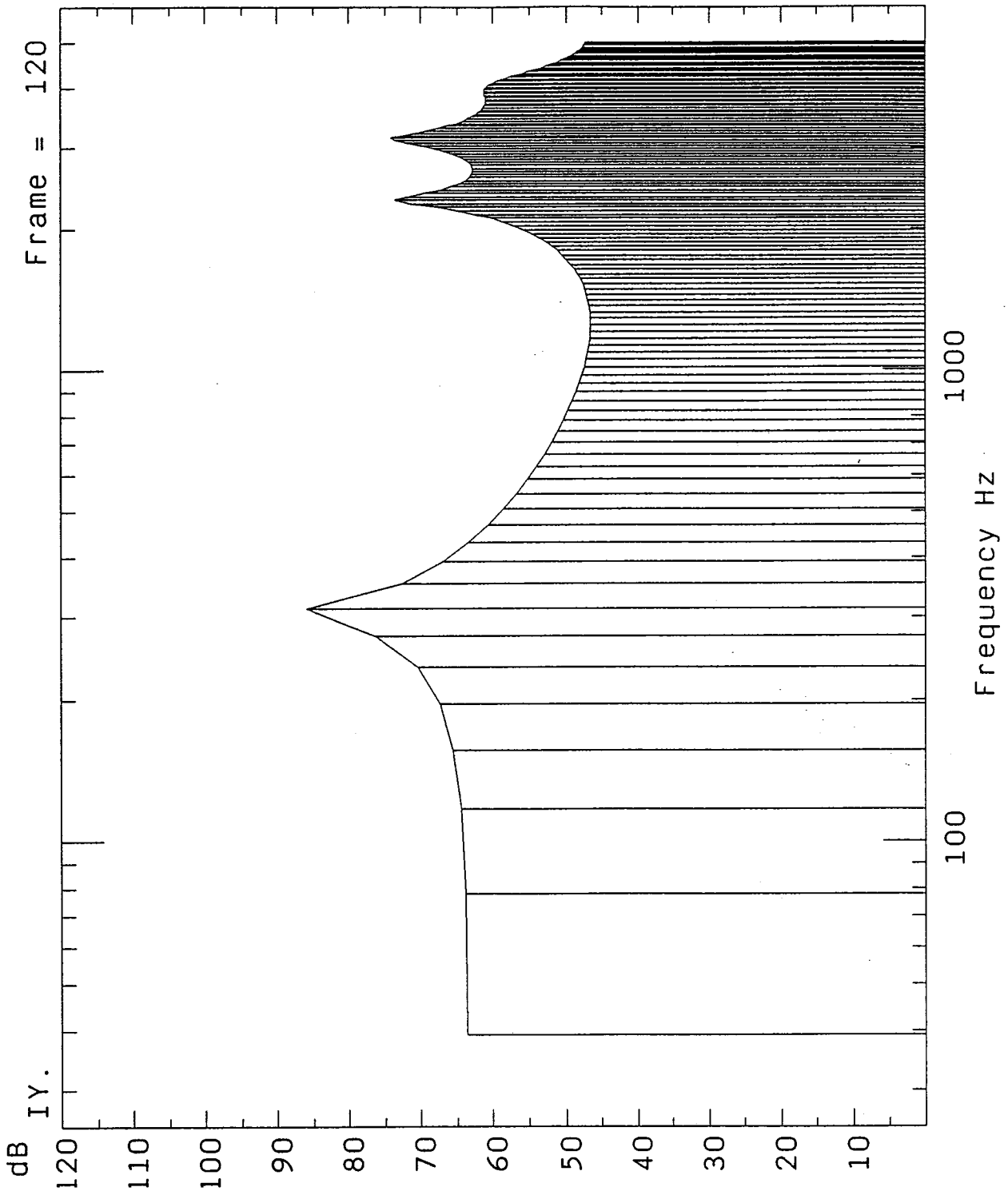


Figure C-2: Spectral envelope derived from FFT of [IH] reference token.

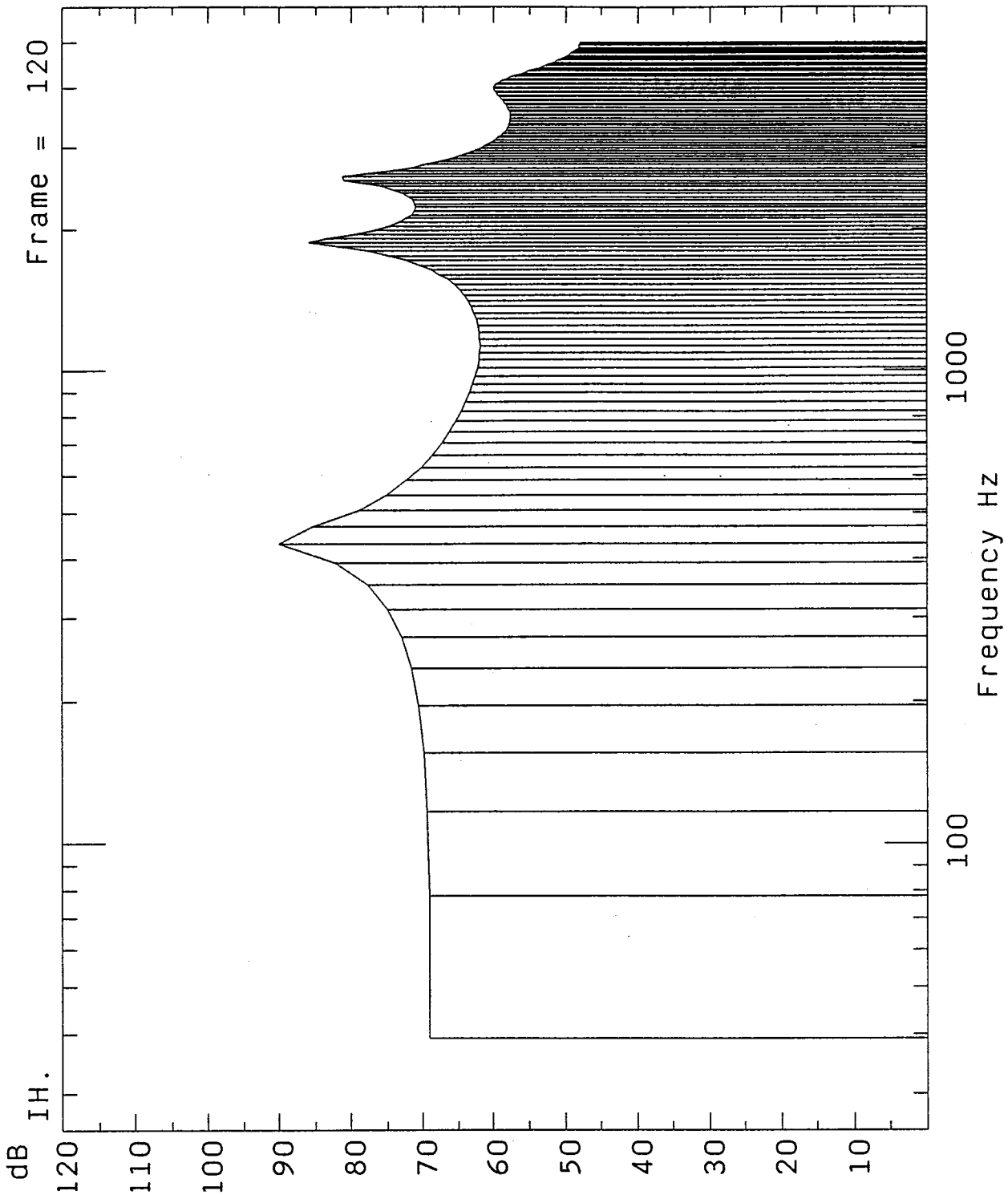


Figure C-3: Spectral envelope derived from FFT of [EH] reference token.

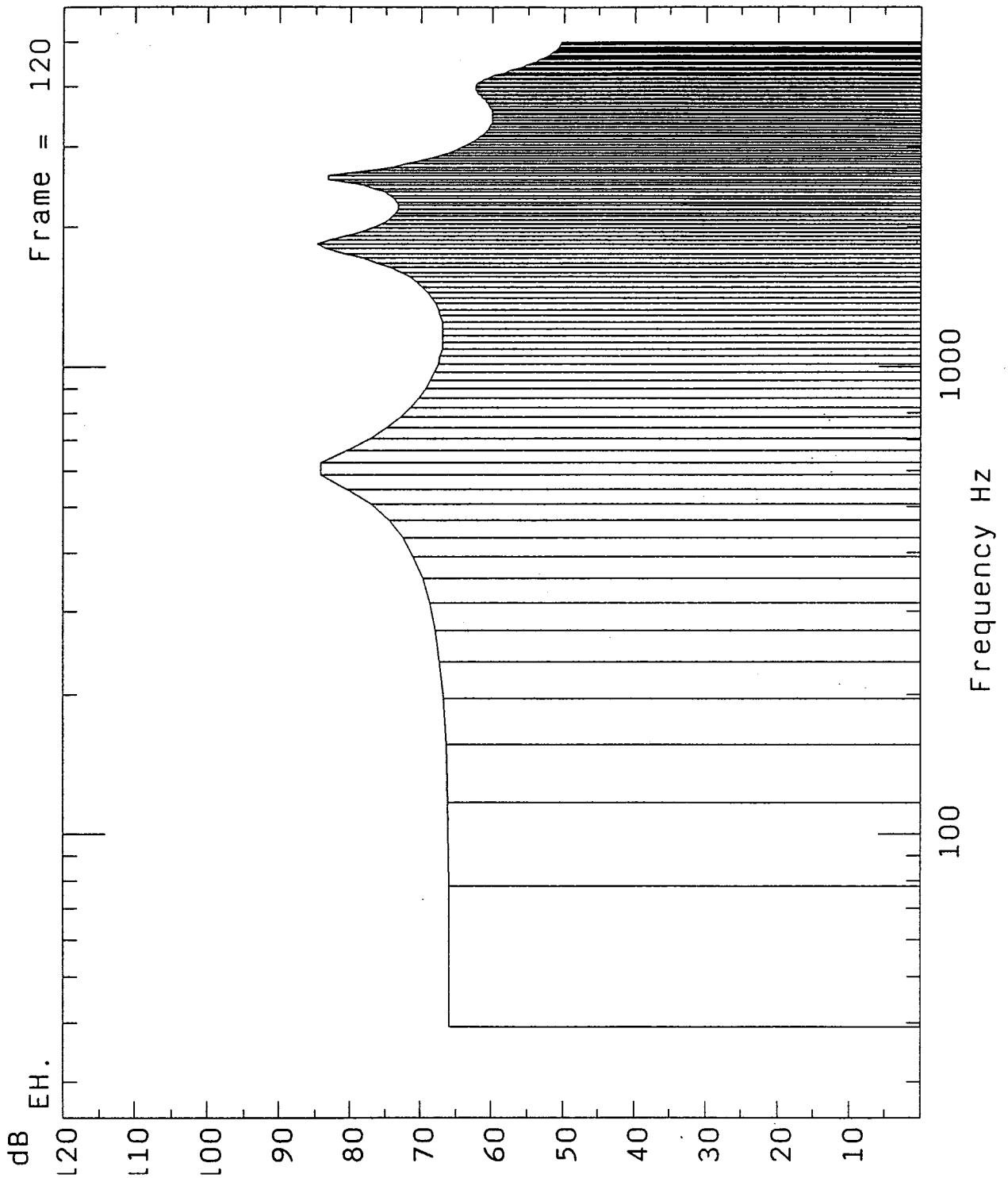


Figure C-4: Spectral envelope derived from FFT of [AE] reference token.

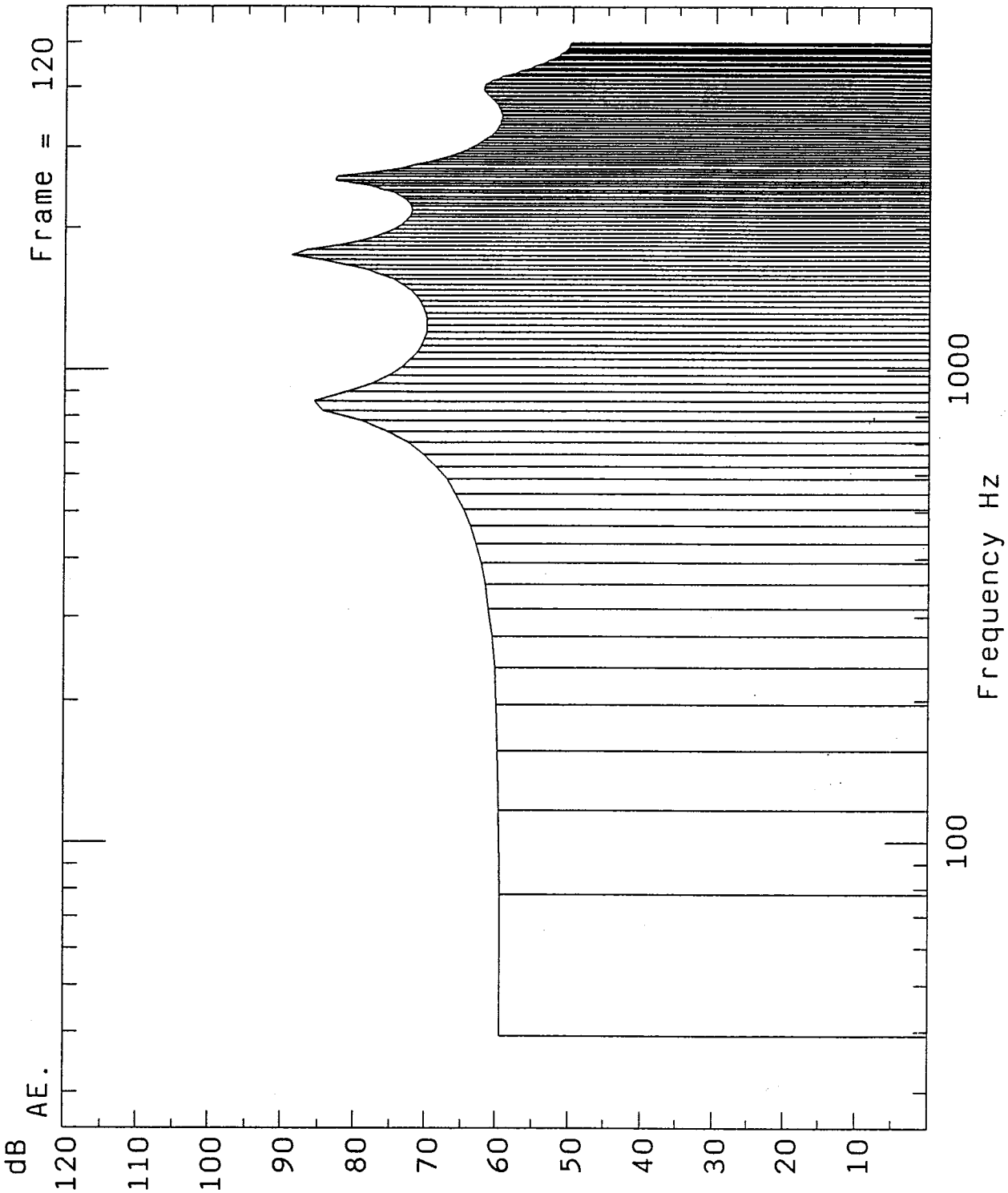


Figure C-5: Spectral envelope derived from FFT of [AA] reference token.

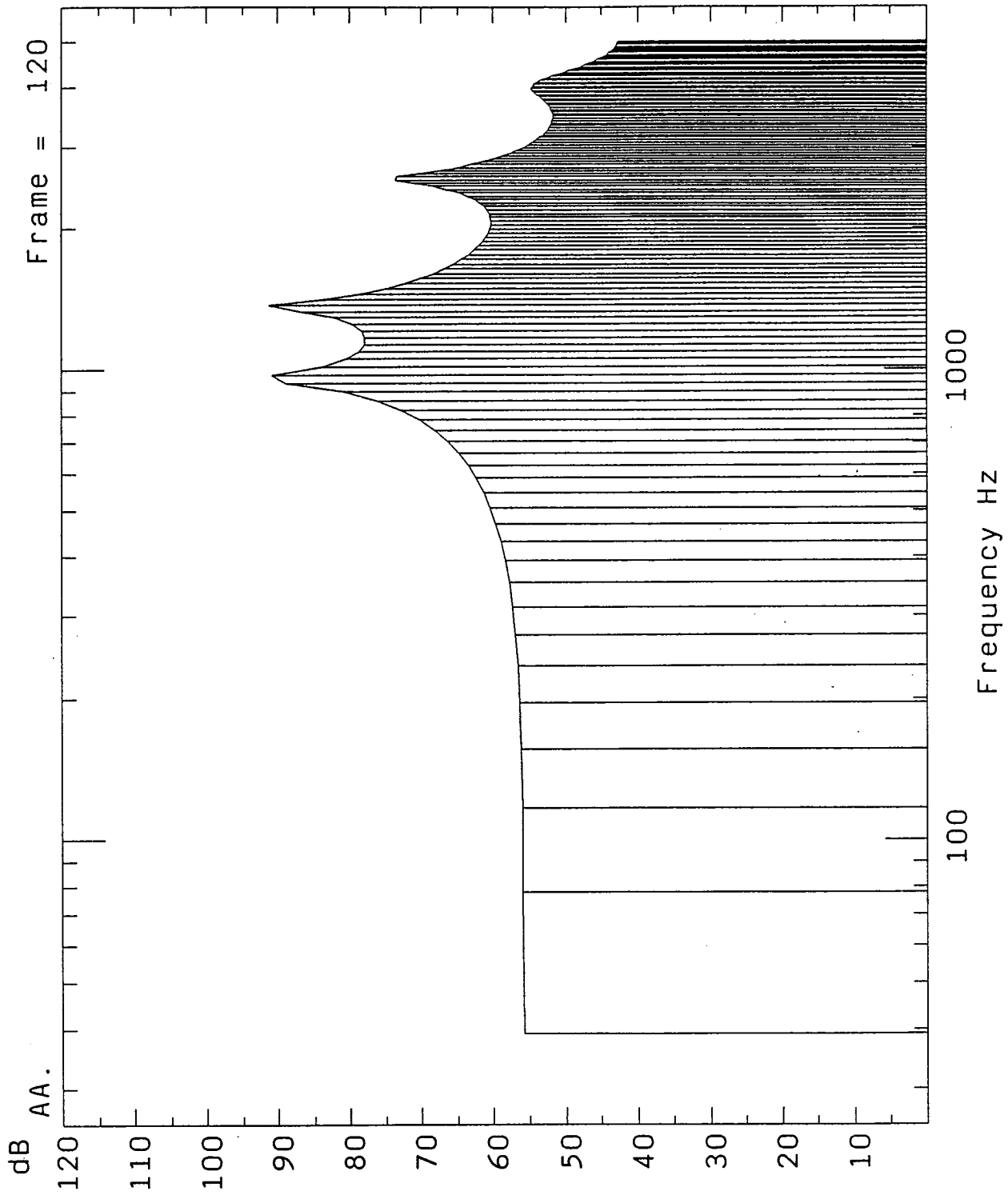


Figure C-6: Spectral envelope derived from FFT of [AO] reference token.

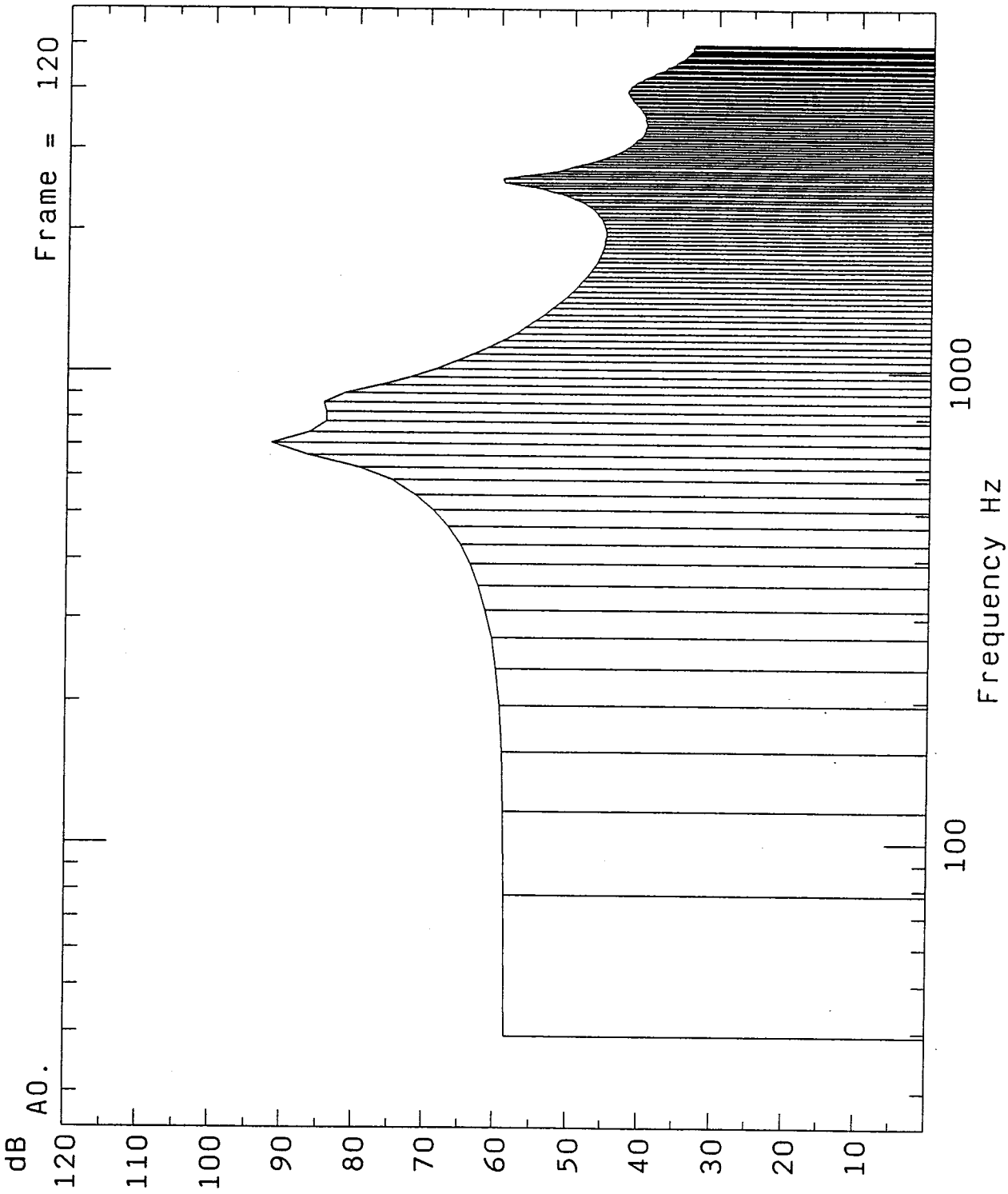


Figure C-7: Spectral envelope derived from FFT of [AH] reference token.

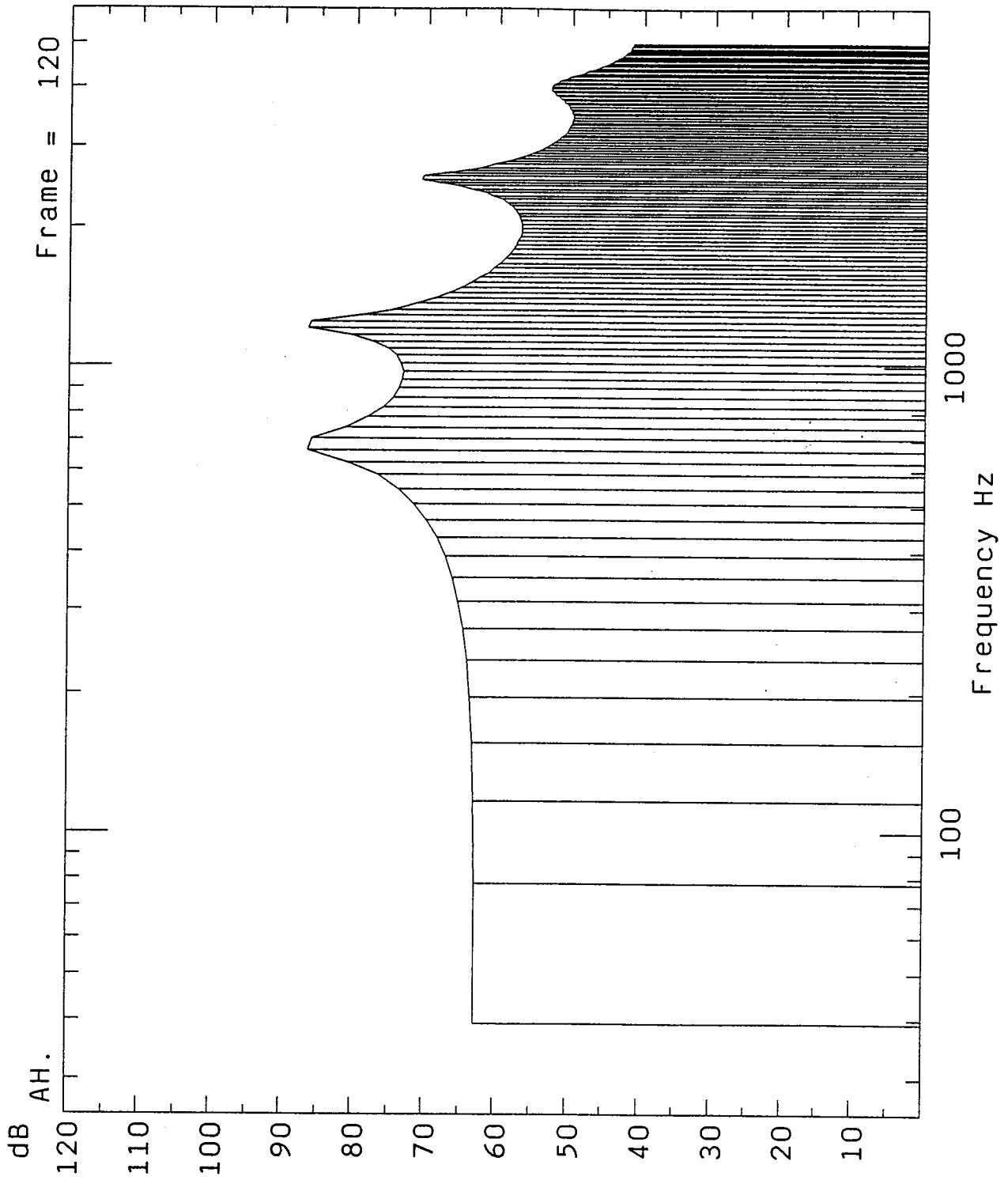


Figure C-8: Spectral envelope derived from FFT of [UH] reference token.

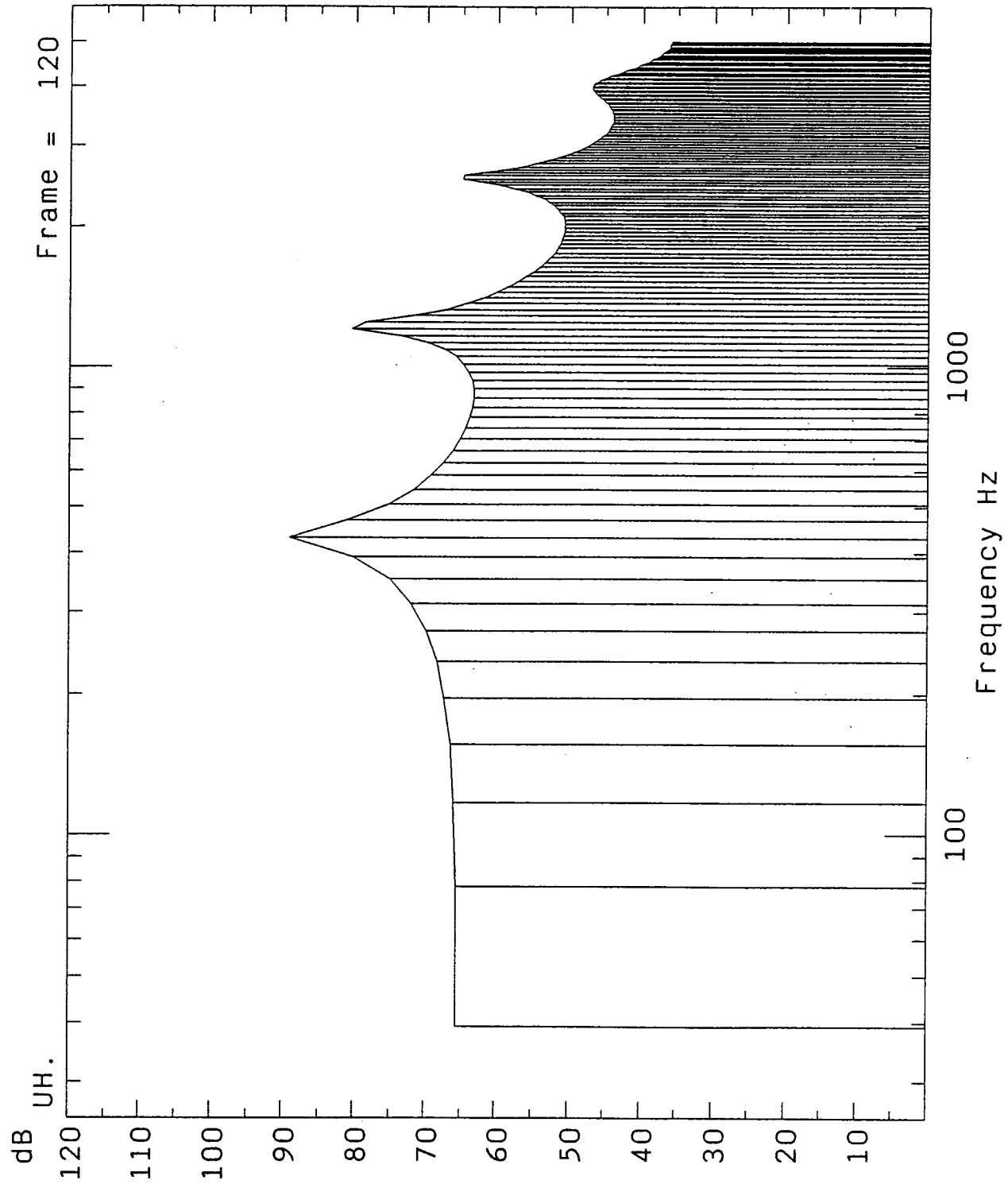


Figure C-9: Spectral envelope derived from FFT of [UW] reference token.

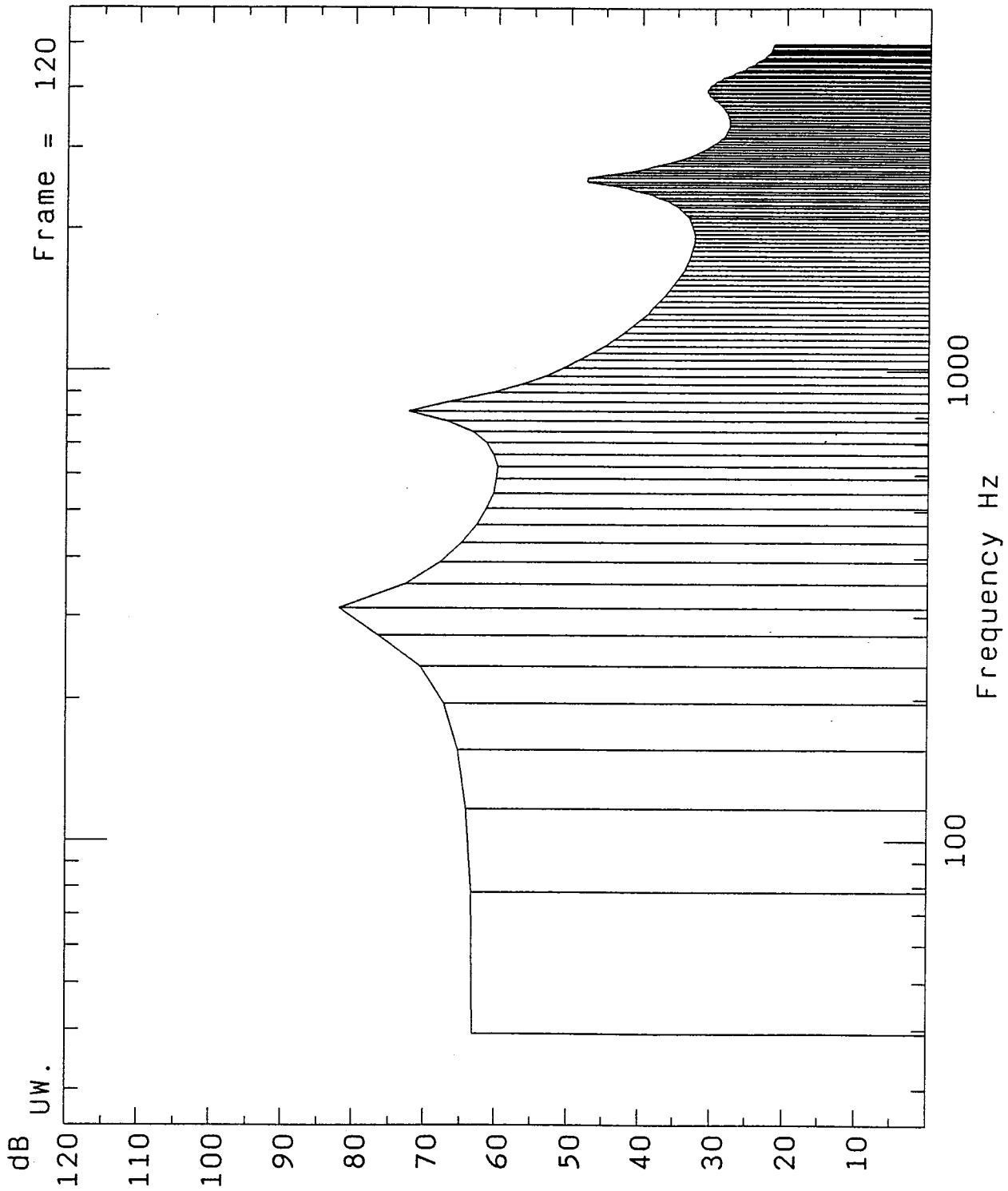


Figure C-10: Spectral envelope derived from FFT of [ER] reference token.

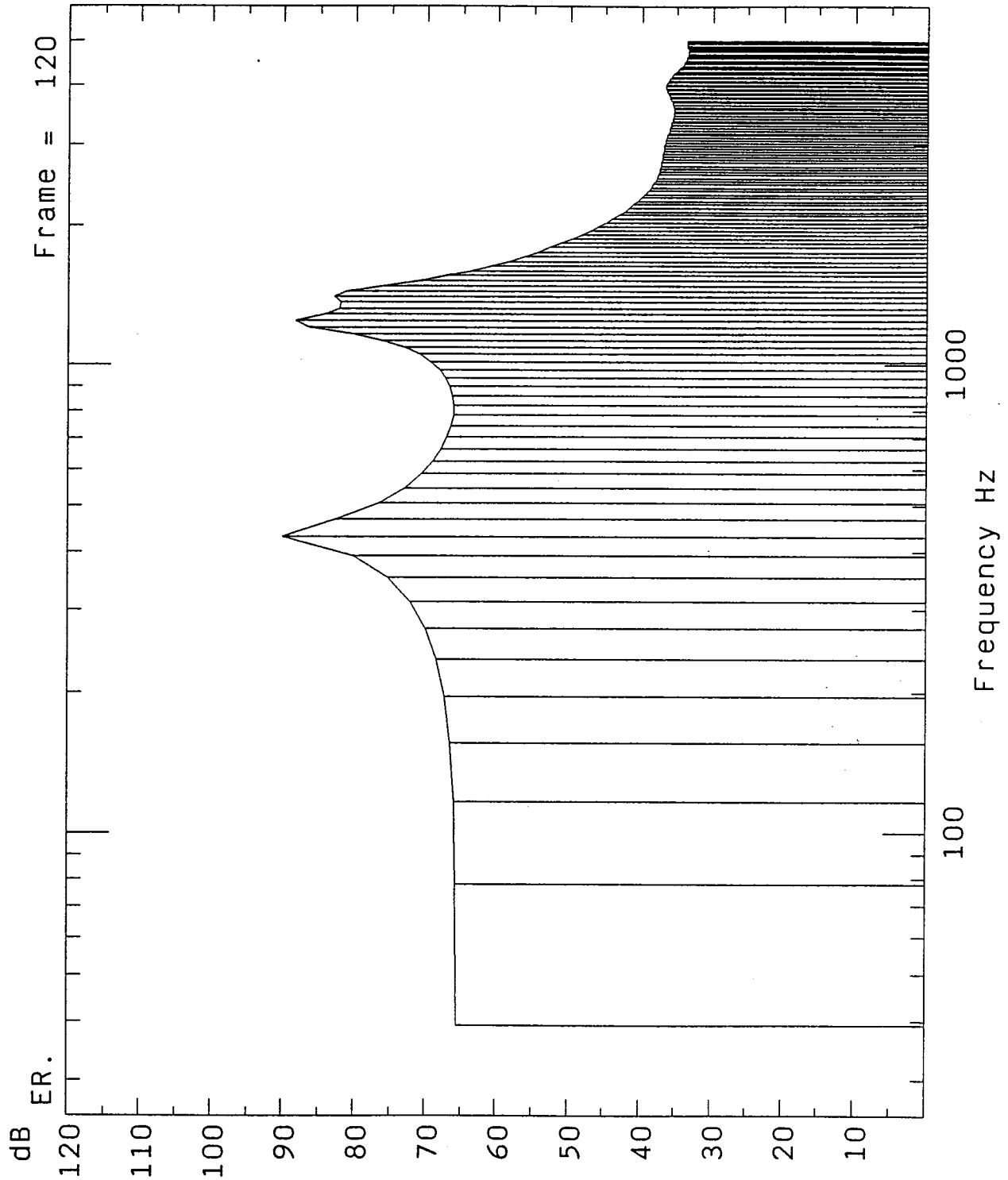


Figure C-11: Spectral envelope derived from FFT of [IY-IH] reference token.

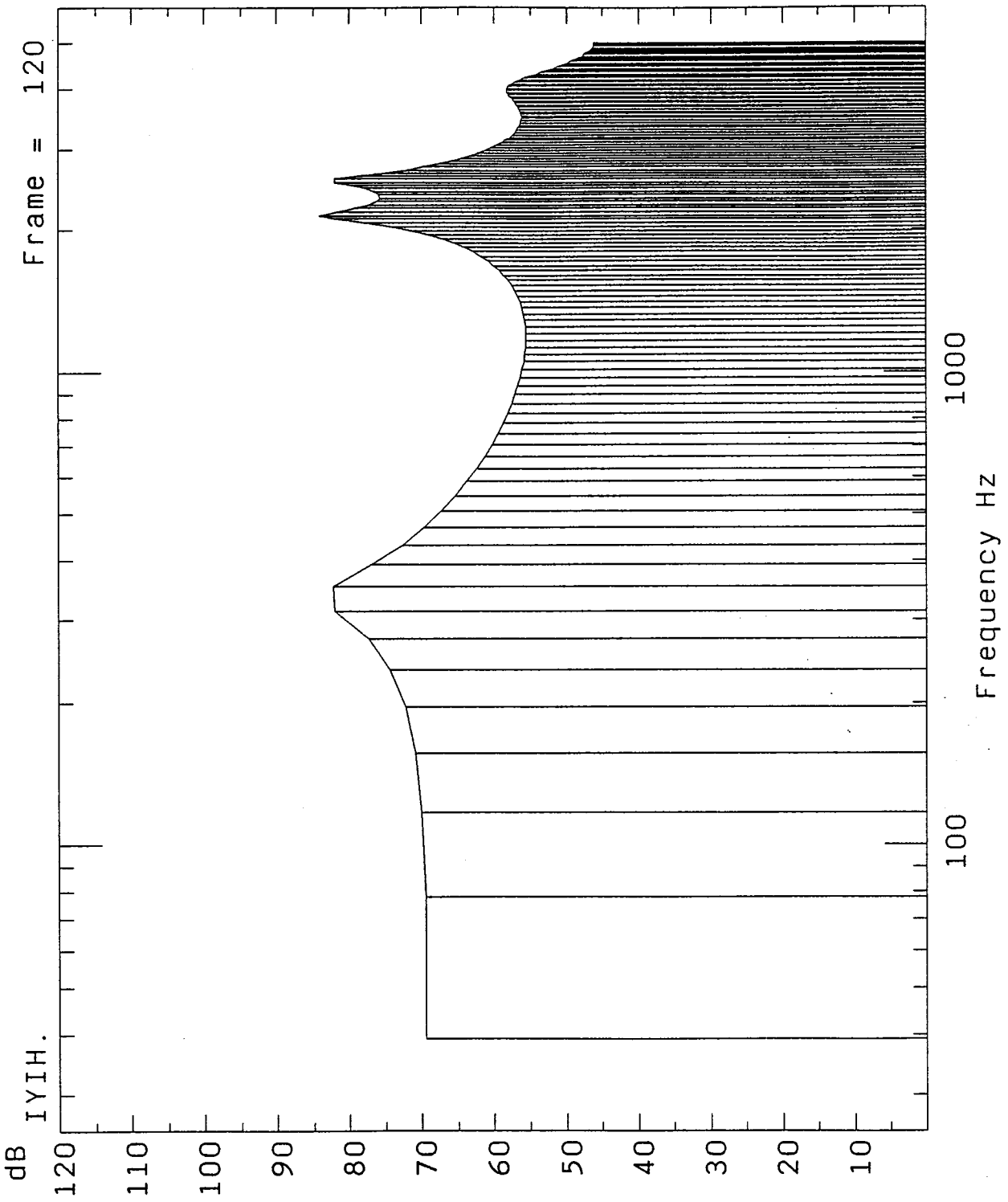


Figure C-12: Spectral envelope derived from FFT of [IH-EH] reference token.

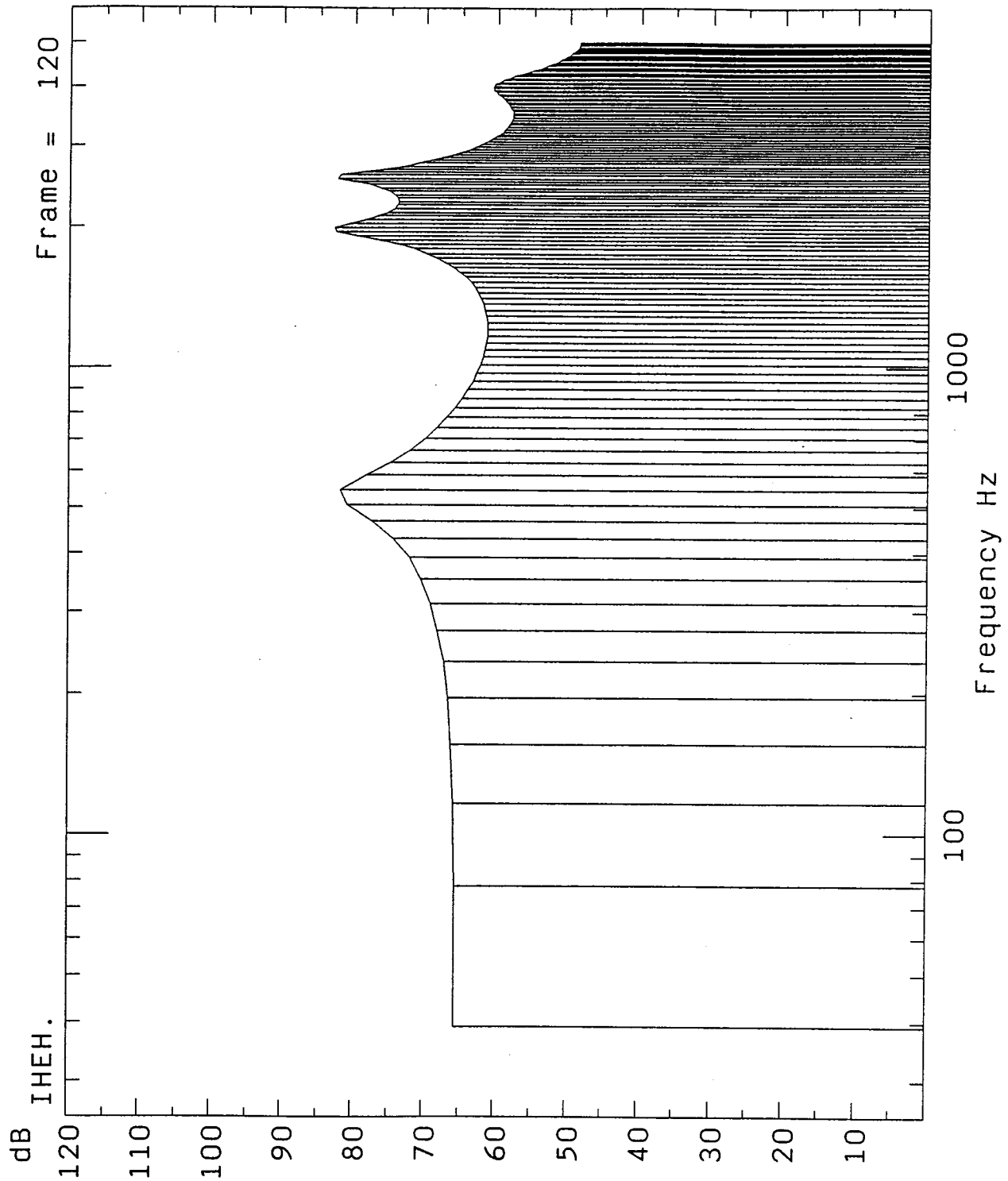


Figure C-13: Spectral envelope derived from FFT of [EH-AE] reference token.

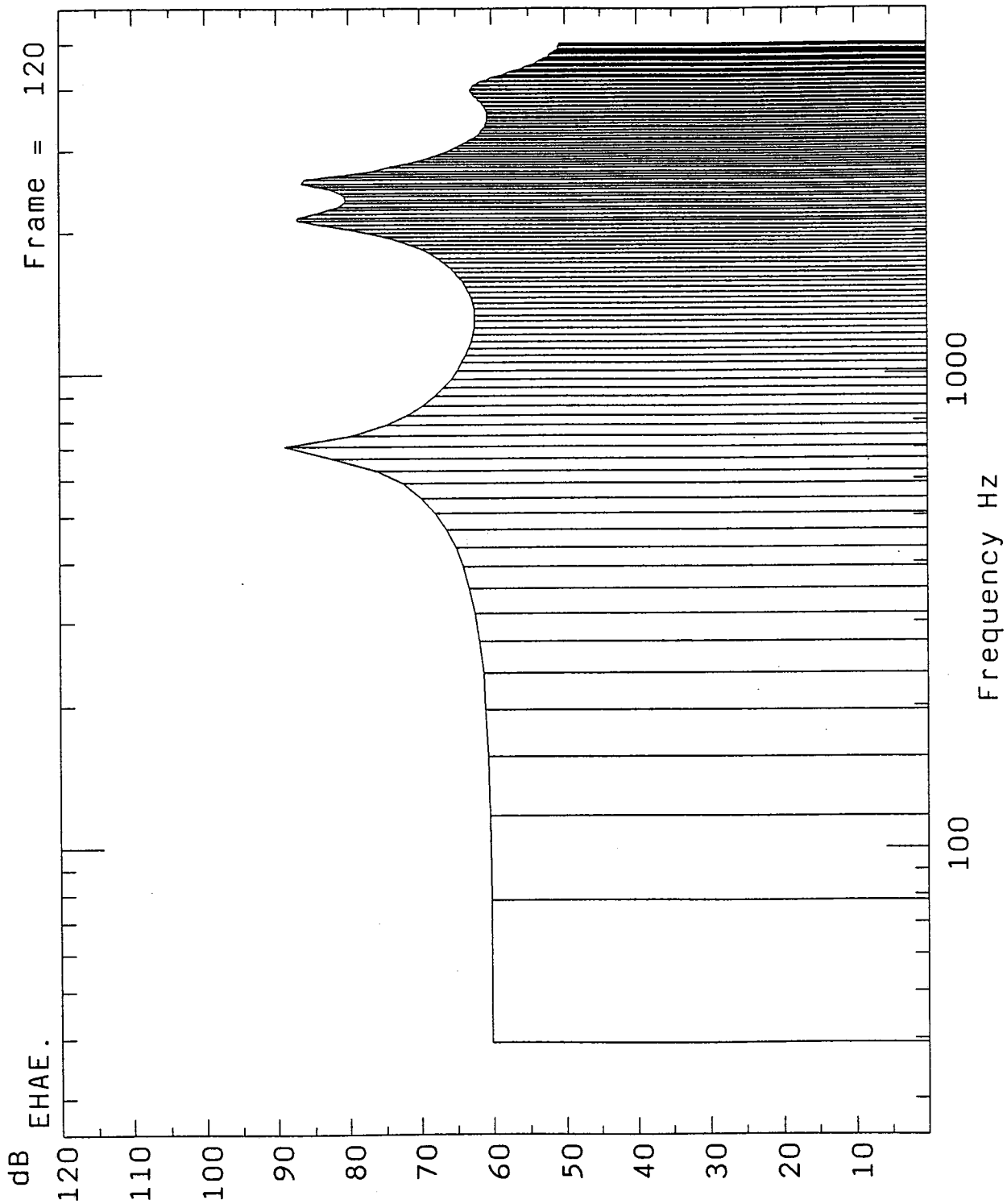


Figure C-14: Spectral envelope derived from FFT of [AE-AH] reference token.

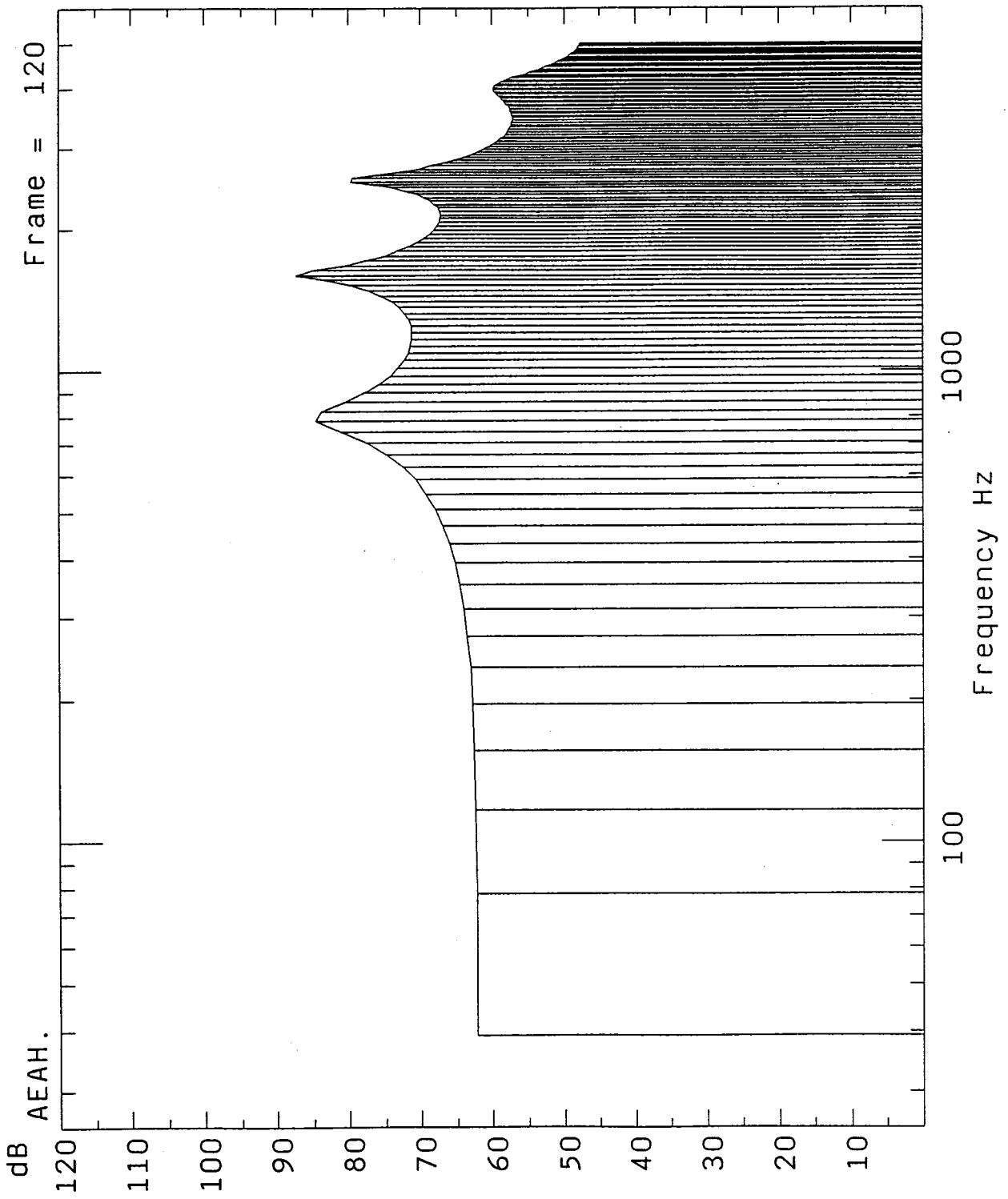


Figure C-15: Spectral envelope derived from FFT of [AH-AA] reference token.

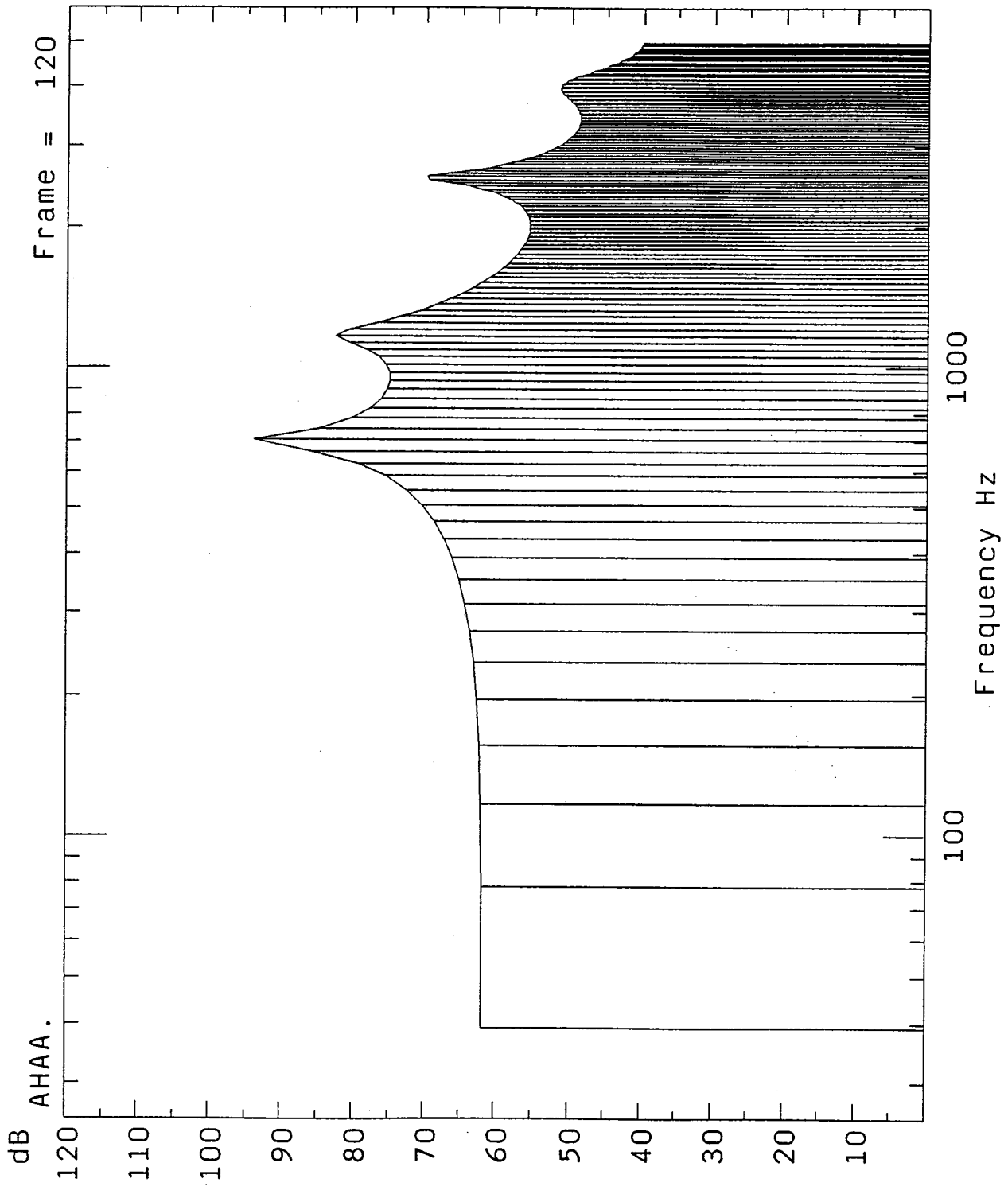


Figure C-16: Spectral envelope derived from FFT of [AH-UH] reference token.

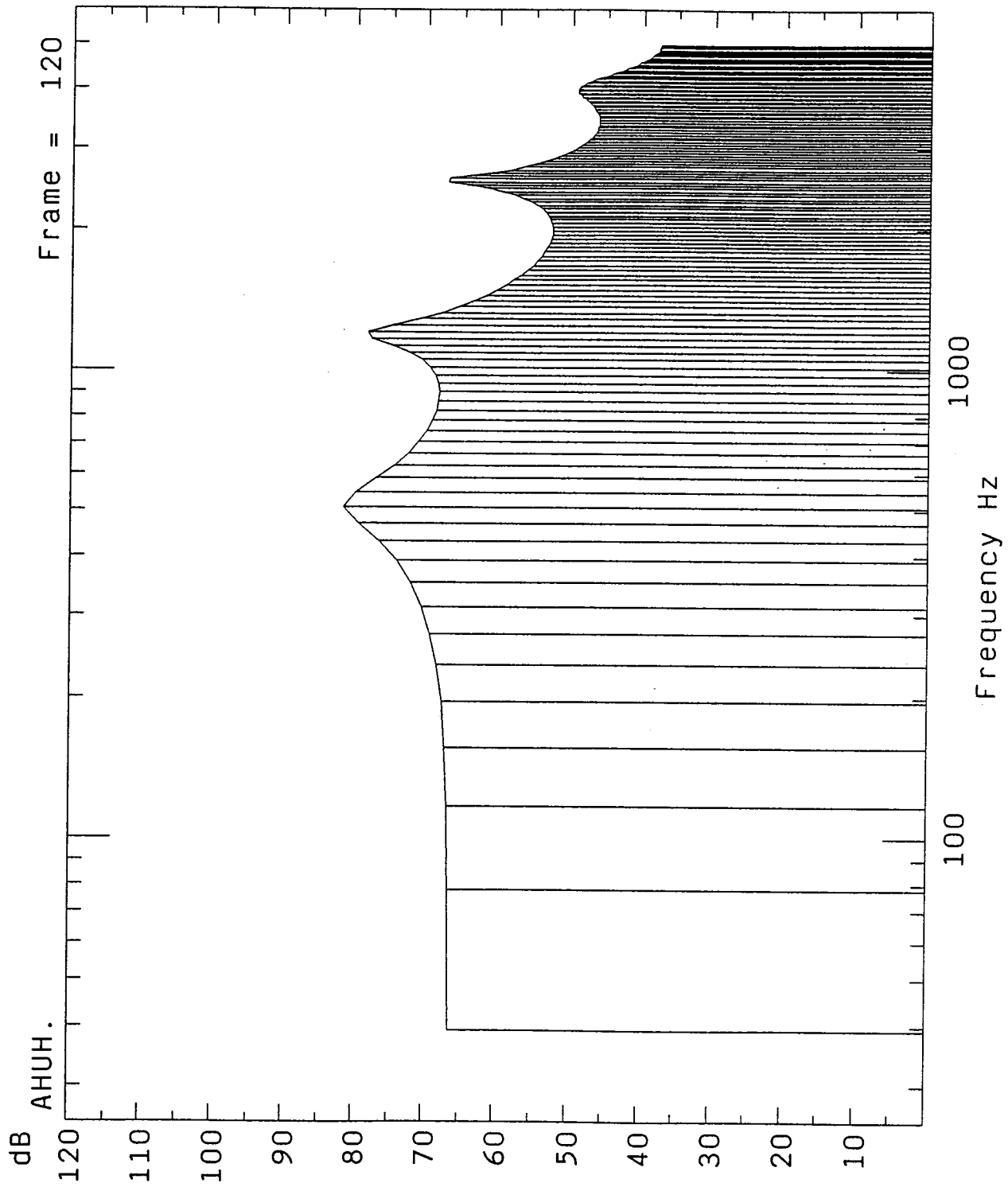


Figure C-17: Spectral envelope derived from FFT of [UH-UW] reference token.